大数据分析与数据挖掘

简祯富　许嘉裕／编著



清华大学出版社

大数据分析与数据挖掘

简祯富　许嘉裕　编著

清华大学出版社
北　京

内 容 简 介

随着移动通信和行动装置普及、物联网和网络发展，以及云端技术的不断进步，现今数据产生、搜集和储存方式比以往更为方便。数据挖掘与大数据分析可以从海量数据中，找到值得参考的样型或规则，转换成有价值的信息、洞察或知识，创造更多新价值。

本书主要介绍数据挖掘与大数据分析的理论方法与实践应用，并加入丰富的实务案例介绍，具体说明如何应用数据挖掘与大数据分析技术以解决真实问题，深入浅出地剖析从数据中掏金的秘诀。全书共分为13章，内容涵盖数据挖掘基本概念与数据准备、数据挖掘的方法与实证、数据挖掘的进阶运用；书中也提供R语言与编程实例辅以说明，使读者更能融会贯通地应用数据挖掘方法，进而提升大数据分析和数字决策能力。

图书在版编目(CIP)数据

大数据分析与数据挖掘/简祯富，许嘉裕编著. --北京：清华大学出版社，2016
ISBN 978-7-302-42425-3

Ⅰ. ①大…　Ⅱ. ①简…　②许…　Ⅲ. ① 统计数据－统计分析　②数据采集　Ⅳ. ①O212.1　②TP274

中国版本图书馆 CIP 数据核字(2015)第 306758 号

责任编辑：冯　昕
封面设计：张京京
责任校对：王淑云
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　**邮　　编：**100084
社 总 机：010-62770175　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　**印　张：**23　**字　　数：**560千字
版　　次：2016年3月第1版　**印　　次：**2016年3月第1次印刷
印　　数：1～2000
定　　价：49.00元

产品编号：065663-01



前言

Foreword

“为大于其细，行远必自迩！”

1992年我到美国威斯康星大学麦迪逊分校（UW-Madison）攻读决策科学与作业研究博士时，发现我在新竹“清华大学”念的概率、统计、实验设计和统计方法等课程的教科书作者竟然都是麦迪逊的教授，所以选择统计作为副修；另一方面，我又在麦迪逊的医疗系统研究分析中心（Center for Health Systems Research and Analysis，CHSRA）担任研究助理，参与由Gustafson教授领导的大型研究团队发展的“综合医疗促进支持系统”（Comprehensive Health Enhancement Support System，CHESS），计划的目的是借着提供信息（information）、转介服务（referral to service providers）、决策支持（decision support）和社会援助（social support）等方式，帮助面对疾病和健康危机的人（如癌症和艾滋病患者）及其亲友取得相关信息、寻求可利用的资源、分析决策，以及社群服务和互相扶持等。我的主要工作是分析系统所搜集的使用数据和用户填写的问卷调查数据等，并在每周研究团队的定期会议上进行汇报，通过各种可能的分析和数据探索，以证明CHESS的效益。因为我的指导教授当时只是团队中的助理教授，所以我特别卖力分析，生怕工作不保就没有奖学金了。有一天，研究团队的一位成员在会议后告诉我说，我做的工作好像“数据挖掘”（data mining），他认为数据挖掘的方法将来可能会超越统计，虽然当时我觉得怎么可能有一种最近才发展的方法，可以超越已有几百年根基的统计学，但也让我注意到数据挖掘这个研究领域。

1996年我回到新竹“清华大学”任教，即成立“决策分析研究室”（Decision Analysis Laboratory，DALab），和研究伙伴与学生们包括本书共同作者许嘉裕博士一起投入决策分析、数据挖掘和优化的研究和实践工作，并通过产学合作计划作研究，然而却苦无合适的教材训练学生，特别是结合实际案例的课本，因此就持续借着整理产学合作研究成果、撰写期刊论文和指导学生论文之机，准备撰写教科书的基础材料。数据挖掘和大数据分析是方法论，也是实证推导模式（empirically derived model），因此必须结合方法发展与实证研究以检验研究效度。决策分析研究室研究团队与台积电、旺宏、台达电、联发科、广达电脑、创意电子、晶元光电、采钰、关东鑫林、茂迪、普生、力晶、世界先进等公司建立双赢的产学合作机制，做到学术研究贡献能够接连获奖，而实际效益能够达到合作厂商产业化的要求，作为更深一层理论研究的基础；更有幸从2005年借调台积电三年，实际应用所发展的分析方法在企业营运中，领导研究室的学生们和工业工程处同仁们一起推动台积电“IE十大建设”并发展相关的分析技术和数字决策系统，提供数字化系统化之决策依据，而从中得到产业导师宝贵的指导和回馈，也累积实战的经验和心得；进而执行台湾“科技部”“IC产业同盟”（Semiconductor Technologies Empowerment Partners Consortium，STEP Consortium）暨

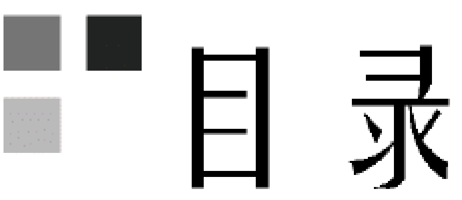
深耕工业基础技术计划，并成立“清华-台积电卓越制造中心”(NTHU-TSMC Center for Manufacturing Excellence)，把累积多年的实证及大数据分析技术，推广到半导体供应链上、下游和其他高科技产业，借此提升产业的决策分析和智能制造能力；并通过主办“清华IC学堂”“半导体大数据分析竞赛”及产学合作成果发表研讨会等活动，培养具备跨界创新、团队合作能力的“资料科学家”。因此，本书在编撰过程中一再修改更新，希望一方面能深入介绍数据挖掘与大数据分析的基础方法和工具，另一方面则通过跨领域的实际案例和范例程序，以具体培养结合理论与实务的决策科学家。

非常感谢新竹“清华大学”和元智大学的良好学术研究环境和科学园区的地利人和，使我们可以结合理论与实务，从产业大数据和具体问题的实证中发展适用的方法、检验所学，再进而导向更深一层的研究。随着问题的广度和复杂度以及合作伙伴的阶层和领域而不断成长，这一路走来，虽然整个研究团队一直秉持自强不息、行胜于言的精神努力提升，但也得力于产业先进和合作伙伴们的提携协助和计划执行过程中的指导，因此要感谢的人非常多，希望借着本书的出版能使更多读者从中得到启发和实际的帮助，以造福社会和产业，也算是间接回报所有关心和帮助我们的人。尽管本书经过长期的准备，但完稿阶段所花费的心力远远超过预期，特别感谢专任助理梁婉玲编辑汇总的工作和与出版社的联络，减少本书错误的可能，以及决策分析研究室同学们一起打拼完成各项研究计划，这也是本书各案例的论文均引用完整作者名单的原因；也感谢在“数据挖掘”课程教学中每位互动的学生，让我们得到教学相长和调整教材的回馈建议。本书自2014年在台湾出版以来，引发学术界和产业界的广泛回响，成为多所大学和各大企业的指定教材。感谢北京清华大学出版社理工分社张秋玲社长和冯昕主任的支持，将全书重新编辑改版，去芜存菁，并增添一章全新章节，使内容更加丰富完整。然而，本书疏漏之处在所难免，盼诸位领导和前辈，不吝赐教，以提升大数据分析和数字决策能力。

简祯富　许嘉裕　谨识

IC产业同盟，2015冬





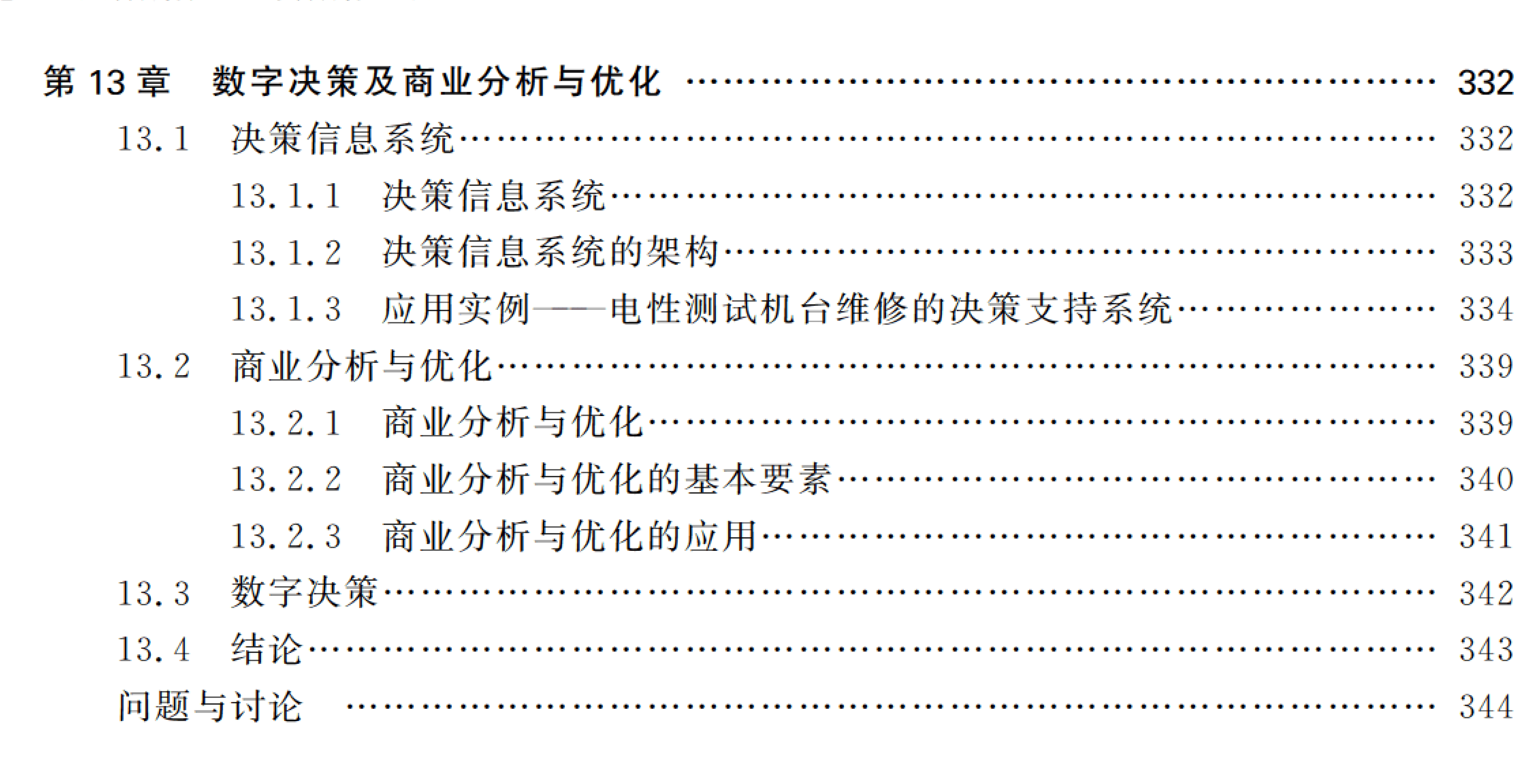
目录

Contents

第 1 篇　大数据分析与数据挖掘导论

第 2 篇 数据挖掘方法与实证

第 3 篇　数据挖掘进阶运用



第 1 篇

大数据分析与数据挖掘导论

第 1 章

大数据分析与数据挖掘概论

1.1 前言

随着信息科技的进步和网络的发达、计算机运算能力的增强以及数据搜集与储存技术持续改进的影响，大幅改变数据的分析和应用方式，**“大数据分析”(big data analytics)**和**数据挖掘(data mining)**可以发掘先前未知且潜在有用的信息样型(patterns)或规则(rules)，进而转化为有价值的信息或知识，帮助决策者迅速做出适当的决策，是现代企业重要的竞争优势。

由于自动化的生产环境、智能手机的普及、电子商务的发展、物联网的建立以及社交网络的发达，现在多数人都可以不受时空地点限制地上网，浏览社交网络，在网络上聊天、购物，以及实时收看与查询最新的新闻报道与文章等，也可以用来管理远程的生产和服务系统。当你在微博上打卡点赞、收发电子邮件、到便利商店购买零食、搭乘大众交通工具、经过停车场利用信用卡缴费时，这些日常生活中的习惯与动作，随时随地正透过网络记录，快速累积成巨量数据或大数据。过去对商品的评价主要是通过口口相传，而现在则是借由在线文章发表，由社交网络快速扩散，这意味着网络经营的重要性已开始逐渐大过实体经营，大数据分析正引领着数字决策并带来新商机。

“数据”在经济学中属于非竞争性的商品，其与物质性的东西(例如食物、车等)不同，并不会因为使用次数增加而降低价值或造成耗损。因此，零售业者累积的事务数据可以一再使用，根据不同目的提取不同的数据，或运用于不同的目标对象上(Mayer-Schonberger & Cukier，2013)。除了传统的统计分析和数据挖掘外，大数据分析技术和应用正改变我们的生产方式、服务系统和生活形态。

每一秒，一间大型医院会增加 12 万笔健康相关的生理数据；每一分钟，YouTube 网站会接收到民众上传总长达 72 小时的视频；每一天，一家银行的信用卡交易次数达 500 万笔。时间分秒走过的同时，大量数据也随时都在快速累积，如图 1.1 所示。而在全世界数兆个传感器、超过五亿部智能手机、十亿台计算机上，每一天不断运作所产生的数据量估计高达 25 亿 GB(胡世忠，2013)。科技研究公司 IDC 更预估，到 2020 年全球数据量将累积达 40 000 ZB(Gantz & Reinsel，2012)，数据储存单位如表 1.1。

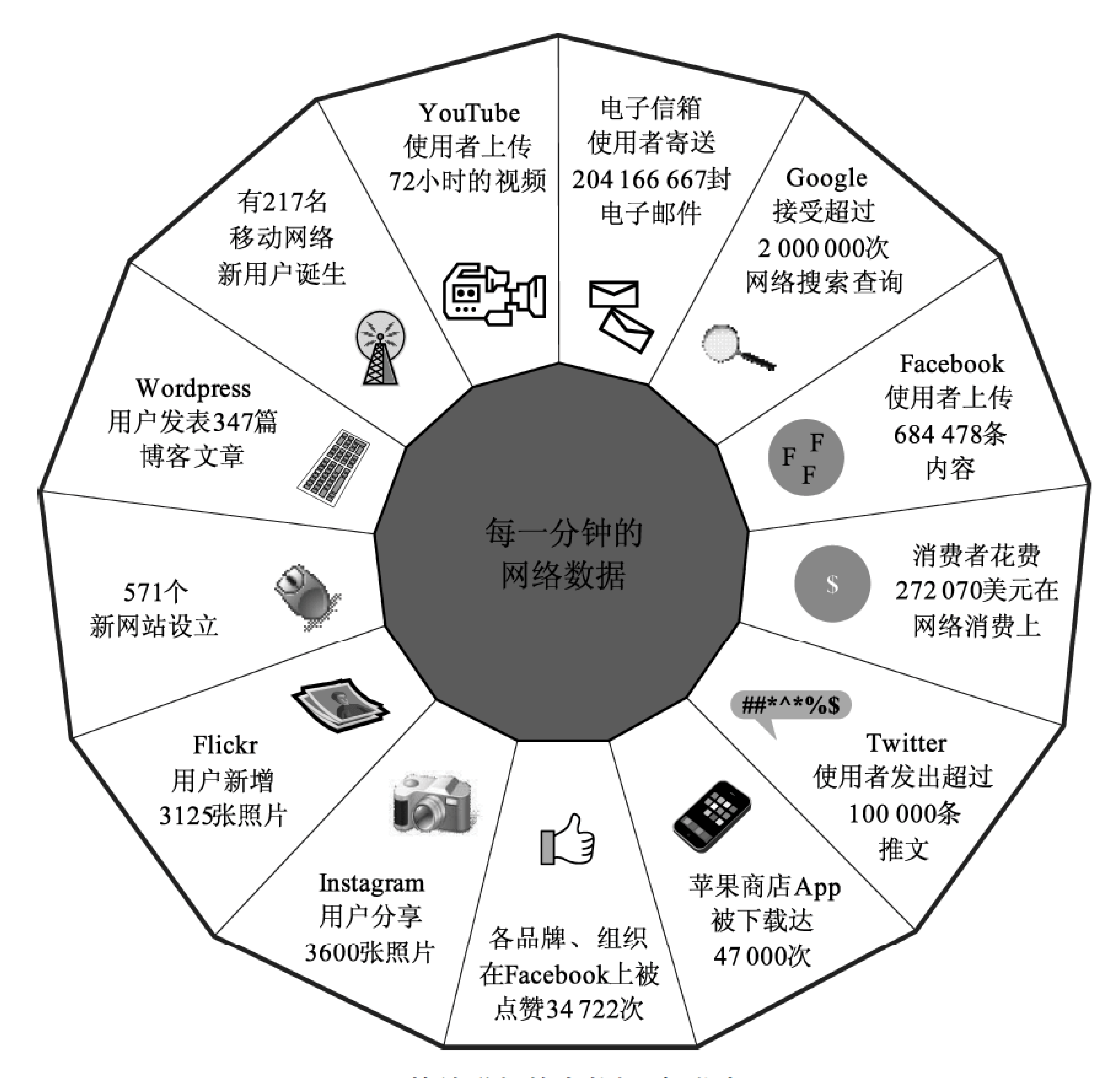


图 1.1 持续增加的大数据(胡世忠,2013)

表 1.1 数据的储存单位

储存单位/B	文件储存单位
Kilobyte (KB)	1 KB=1024 B=2^{10} B
Megabyte (MB)	1 MB=1024 KB=2^{20} B
Gigabyte (GB)	1 GB=1024 MB=2^{30} B
Terabyte (TB)	1 TB=1024 GB=2^{40} B
Perabyte (PB)	1 PB=1024 TB=2^{50} B
Exabyte (EB)	1 EB=1024 PB=2^{60} B
Zettebyte (ZB)	1 ZB=1024 EB=2^{70} B
Yottabyte (YB)	1 YB=1024 ZB=2^{80} B

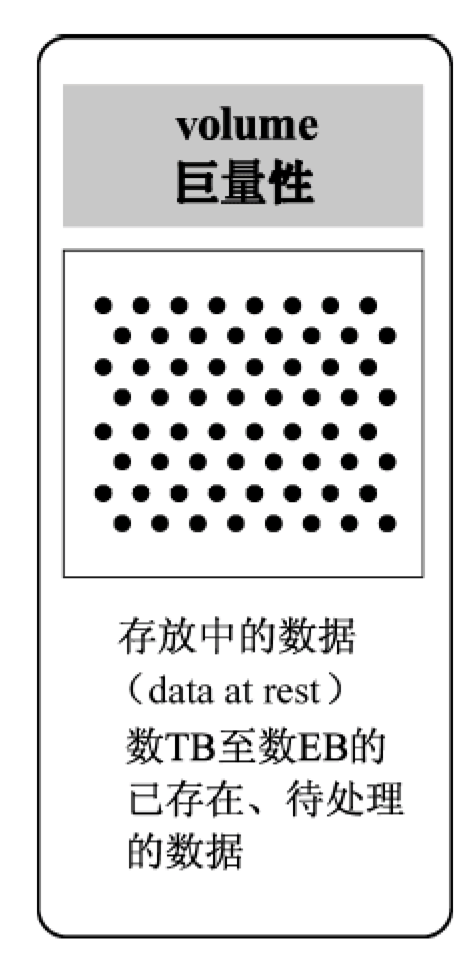
大量的传感器与电子卷标置入到日常生活的电子设备中,例如手机、监控摄影机、环境温度传感器、水电天然气表等,随时感测人们的生活动态。例如,电力公司为了节省能源,开发的智能电表和智能电网即装置了大量的传感器,24 小时不间断地测量与传输终端顾客的电力使用信息。对终端顾客而言,智能电表能实时显示家中的用电量,协助用户调整用电习惯。对电力公司而言,则可透过实时用电量的监控,掌握电网供电状态,当耗电量可能超过

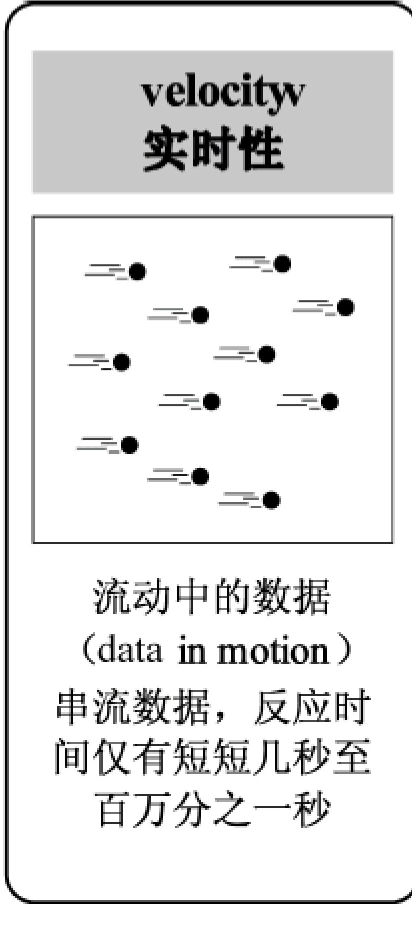
标准时，尽早采取备用措施，降低可能的无预警停电。

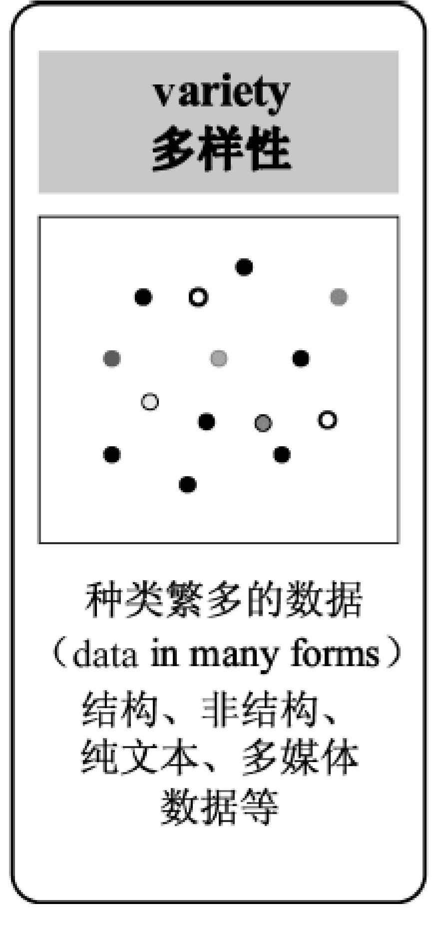
大数据的运用，首先，必须厘清客观记录的数据、分析处理所得的信息以及了解从信息所衍生出的知识之间的差异。

“数据”(data)是对事件审慎、客观的记录，而记录的目的在于创造信息的重要原料。数据是以一种结构化的方式记录事件发生的相关数据，例如，零售店 POS(point of sale)系统的一笔交易项目、时间与金额，医院药店的一笔就诊与给药记录，以及半导体厂生产在线工件进入与离开某加工设备的时间等。大数据一般来说是无法被传统工具直接处理、分析的数据，大多是半结构化以及非结构化数据，仅有少量是结构化数据。结构化数据指的是有关联性定义的固定结构数据，例如数据库中的每一笔数据都要按照事先定义的格式与顺序储存，否则无法被读取；半结构化数据则具有一定程度的编码设定与格式，但仍有部分数据无法统一格式，例如电子邮件、XML、HTML 的网页数据；非结构化数据则没有统一格式，例如图片、声音、影像等数据。

大数据一般具有 4V 特性(如图 1.2 所示)：①volume，其代表的不仅是庞大的数据量，更重要的是“母体”数据的完整性，因此，不像过去多以统计来处理少量的样本数据，在“样本＝母体”的趋势下，对于数据的分析与处理也必须发展出崭新的做法；②velocity，数据变动速度快或实时性说明的不只是数据产生的速度快，亦表示系统与分析者也需快速进行分析与反应；③variety，数据多样性则是说明大数据的多元数据种类，如电子邮件、文字、图片等非结构化或半结构化数据；④veracity，数据真实性或不确定性，表示当数据源更加多元且复杂时，数据本身的精确性、置信度及质量也需要经过适当的检验(胡世忠，2013；Mayer-Schonberger & Cukier，2013；Schroeck *et al*.，2012)。







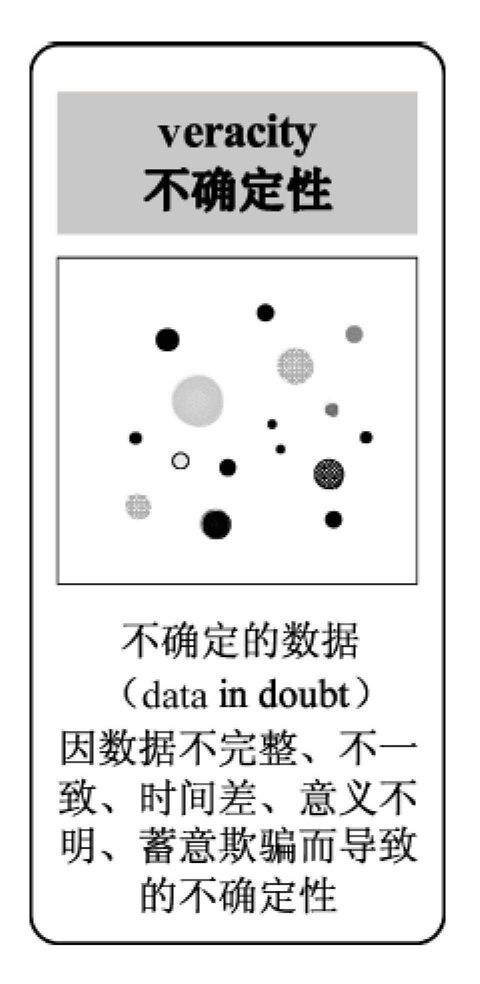


图 1.2　大数据的 4V 特性(Schroeck *et al*.，2012)

因此，随着大数据时代来临，所涉及的数据量规模和数据挖掘的复杂度已经大到难以用简单的方法在合理时间内分析整理成为有用的信息。

“信息”(information)是数据经过处理并赋予意义后，进而转变成具有潜在价值的分析结果。信息可以影响接收者的想法和判断，且具有关联性和目标，通常通过文件或网络，在组织内传送流动。例如，零售店每月交易金额最高的十项商品、每日交易的高峰时段、医院库存药品中超过一季未使用的项目分析以及半导体厂生产在线加工设备的利用率

(utilization)等。

需特别注意的是，数据转换为信息时，转换的“质量”比转换的“工具”更加重要。若是其转换逻辑只是硬把巧合当成规律，则转换得到的信息将不足以用来协助决策。

“知识”(knowledge)来自信息，但并不仅止于信息所传递的信息。它综合了经验、价值及信息，并且成为一种接收、评估、整合其他新经验的架构，例如，专家提出的洞见等。知识存在于文件和储存系统中，也遍及在日常工作等规范中。克尔(Kerr，1991)认为知识应用的重点在于如何阐释数据的意义，彼得·德鲁克(Peter Drucker)亦认为，在数据转换过程中，经理人需要相关知识才能提升转换的信息质量；换言之，信息转换到知识的所有环节都需要“人”的参与。

以《论语》为例，论语里的每一个“字”，若一个个分开来看，代表的是特定意思的单字，就像客观的记录，可以将它视为资料或数据。如果将数个单字合起来就成为“词”或“句”，每一个词句就好像信息一般，都有它独特而可以被发掘诠释的含义。若将更多词句组合成文章诗词的章节段落或是一整本书，配合读者个人的经验、思考和诠释应用，则可用以抒发离骚激起共鸣，也可以传递前人治国平天下的思想结晶，这些信息的组合和相关的思考应用过程称为知识(图 1.3)。从数据、信息到知识与决策，是一连串加值的过程。

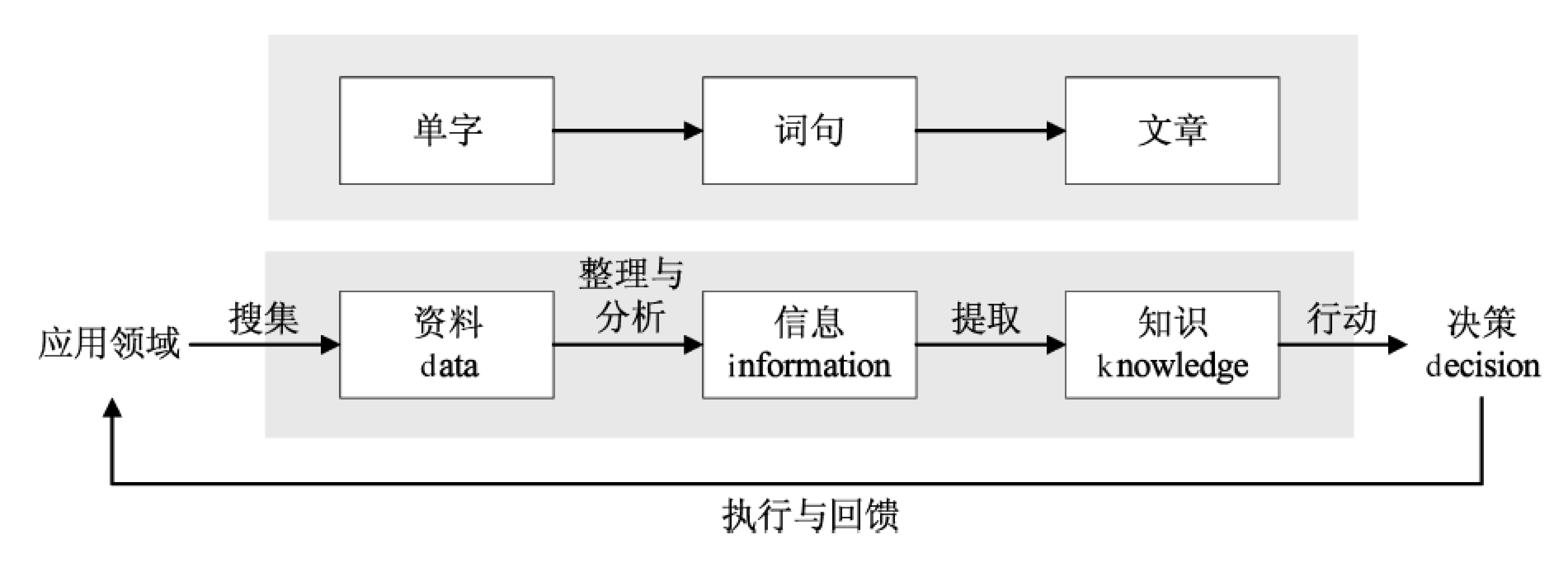


图 1.3 数据、信息、知识的转换与决策

以数据挖掘和大数据分析为基础的数字决策是**“探索驱动”(discovery-driven)**，而非**“假说驱动”(hypothesis-driven)**。以著名的“啤酒与尿布”为例说明两者的差别：探索驱动的分析中，分析人员一开始仅先设定分析方向，如研究消费者购买行为，对于从数据中会找出什么信息并无默认立场，经由数据挖掘的过程发觉数据间存在“啤酒与尿布出现在单次交易记录的频率以星期五晚上为最高”的样型后，深入分析其含义，才成为一个对管理者有效的信息和决策依据；反之，假说驱动的分析则是先假设消费者于星期五晚上购买啤酒后也会顺便买尿布，再进行统计验证此假设。然而，在无相关经验前，一般人不会先提出这样的假设来检验。换言之，通过探索驱动的分析和强大的计算能力可以快速处理巨量数据以找出先前未知、但却具有潜在应用价值的信息，可以促进组织成员间的知识流通与互动，增加企业竞争力。

1.2 大数据分析的应用

理论上，拥有更多的数据，代表获得背后的数据价值机会越大，实际上却不然，原因是不同产业所产生的数据类型也不尽相同。大数据应用范围从营销零售、制造生产到政府部门，

以下说明几个大数据分析的应用范例。

伦敦长期的交通堵塞众所周知，为了配合 2012 年伦敦奥运会期间涌进的 900 万人潮，伦敦市政府通过交通监视系统、摄影机以及公共汽车站与地铁站所发送的信息，提供中央交通控管室纵览整个市中心的交通状况，也能够在各种情境下有效率地调度交通工具。伦敦在城市里四处安装高灵敏度 CCD 相机，借由图形辨识系统以监测出哪些地区出现交通拥塞，并依此结果实时调节交通信号的配时长度。另外，通过将各项运动赛事的举办时程、地点、购票人数等数据输入系统之中，可以预测未来伦敦可能涌现的交通拥塞的区域。此外，伦敦政府亦在停车场里安装传感器随时掌控停车位的使用数据，驾驶人在停车场入口处即可从手机的应用程序实时接收到闲置停车位的方位信息，增加停车的便利性。

2009 年的 H1N1 新型流感病毒，混合了禽流感与猪流感病毒，因而迅速地蔓延，世界各国均担心受到感染。早在美国政府发布 H1N1 疫情新闻之前，有几位 Google 工程师早已利用 Google 搜索引擎预测到美国在冬天将爆发流感，并指出可能爆发的州。由于 Google 每天都会收到上亿次的关键词搜索，他们首先选出美国人最常搜索的前五千万个搜索关键词，再比对美国疾病管制局在 2003 年至 2008 年之间的流感传播数据，除了找出可疑的关键词外，Google 更着重分析关键词的搜索频率与地区有无统计上的相关，靠着分析民众在网络上搜索的关键词，找出感染流感的人，不仅可避免延迟的通报，实时的信息更能用于疫情控制，以及避免再次爆发流感(Mayer-Schonberger & Cukier，2013)。

另一个例子是如何寻找目标客户，对孕妇产品零售商而言，找到潜在的具有高消费需求的怀孕妇女极为关键。他们的做法是，先统计过往怀孕妇女的消费历史数据，从数据挖掘中发现这些妇女大约在怀孕三个月后会开始购买许多无香料的乳液，几个月后，再购买营养补充食品，从中建立几项预测怀孕的指标，一旦出现符合预测指针的客户信息，零售商即主动提供相关可能需要的产品列表以及优惠券以刺激消费。

现今网络上随时快速产生来自社交媒体的大量文字、语音、影像数据，例如电子邮件、新闻媒体、社交网站等，了解这些半结构化或非结构化数据的意义并从中提取重要的信息是文本分析(text analysis)与文本挖掘(text mining)的重要任务。一般而言，文本分析的目的主要有信息检索(information retrieval)、文件分群或分类、情绪分析或语意分析(semantics analysis)。随着智能终端设备的普及，民众越来越习惯在网络上分享个人的心情、喜好、信息及评论，企业也开始分析社交网络中的大量文本数据，并试图从中找到消费者对产品的评价与喜好，作为调整营销与产品开发设计的规划。另一方面，社交媒体的分享特性，使重要事件发生时信息量往往会急速激增，因此可作为实时事件分析(real-time event analytics)的监测工具(Pang & Lee，2008)。

大数据分析也被应用于预测设备维修保养，以避免各种机械或设备的重大故障。例如，许多设备如飞机引擎都安装传感器随时记录设备发出的信号，包括温度、震动、压力、流量等，以预防事故发生。一般而言，设备往往不会是突然发生故障，而是随着时间的累积，借由实时分析传感器所搜集的数据，建立监测模型，在发生异常前发出警告，能避免更大的损失。其他相关的应用还包括人体保健预防等。例如，智能手表或穿戴式装置可随时监控病患的血压与心跳，一旦发生异常，系统可立即发送信号给周围的医院，提供病患实时的医疗服务。物联网的应用相当广泛，包括智能电网、智能交通、环境感知等都是未来重要的新兴领域，也是未来巨量数据的来源之一和大数据分析的重要应用。

强化与顾客的关联必须进一步了解顾客，市场细分(market segmentation)即为认识顾客最有效率的途径。市场营销理论的发展，已由以往大量营销，逐渐转变为差异化营销，进一步进入目标营销(target marketing)。由于不同的消费者生活背景不同，其对产品的需求、满足程度、购买动机的要求也不同，使得厂商很难以单一产品满足所有消费者的个别需求，因此厂商必须依据市场需求现况，衡量本身条件，仔细选择某一个或数个目标市场，针对各个目标市场的需求特性设计不同的产品，以达到营销产品的目的。

以英国零售业领导者 TESCO 的做法为例，相较于传统市场以增加市场占有率为主的营销策略，将营销重点着重于投资大量资金与精力在整体市场，以期最大化市场占有率，扩增营业额；TESCO 从过往消费记录的分析发现，其忠诚度最高的前 5%顾客贡献企业 20%的营收，而忠诚度较低的 25%顾客仅贡献 2%的营收。因此，TESCO 找出具有企业获利价值的顾客群，运用前端客户信息搜集这些顾客的消费习惯信息与背景数据，通过购买变量的分析结果，建立顾客偏好模式并制订后端营销策略(Zoratti & Gallagher,2013)，包括：

- **顾客维系**：即保有现有顾客，针对顾客经常购买的产品提供优惠，或依据顾客消费习惯，来决定量贩方式、空间配置和分类原则。
- **顾客活化**：即重新唤醒沉睡的顾客，如针对顾客曾经购买但一段时间未再购买的产品提供优惠。
- **顾客成长**：即增加现有顾客，根据事务数据库得出的偏好模型，制订定制化服务与营销组合方案，搭配更精准的定价，强化营销效能，如采用交叉销售方式，促使顾客购买未曾消费过、但符合其偏好的产品。

此种目标营销方式可节省大量但无从确定成效的营销预算。后续则通过持续地监视并修正原有营销策略，如空间和分类的优化、精准定价以及促销的效能，找出适合不同顾客群的服务与营销组合，例如将精准营销节省下来的成本反映于产品售价回馈消费者，并通过顾客忠诚卡累积点数，以兑换优惠物品或享受店内其他服务，创造正向回流，维持顾客忠诚度与建立长期买卖双方的稳定关系。

1.3 数据挖掘与数字决策

现代企业必须善用信息科技来解决问题、提升效率及提高决策的质量。各阶层的管理人员经常需随时随地做出关乎企业发展存续的重要决策，因此，如何从庞大的数据中，准确、及时并迅速地撷取出有价值的信息，以协助企业经营者迅速做出正确有效的决策，已成为“十倍速时代”中极为重要的议题。

然而，过多的数据也可能成为一种负担。因此，大数据分析与数据价值的创造便成为将数据转换为资产的成功关键。企业所记录或储存的大量数据，对不同阶层的用户亦代表着不同的价值与意义。一般而言，企业数据的管理者与用户可以分为三种层次(Cabena *et al.*,1997)：①数据库管理者(database administrator)，②数据分析者(data analyst)，③企业决策者(decision maker)。数据库管理者接触的数据量最大，但由于未经处理与加值化，其价值也较低；反之，对于企业组织中的管理者甚至决策者来说，借由数据整理而成的信息，以及结合需求所转变的知识，其量虽小，但价值却远胜于未整理过的原始数据，如图 1.4。

企业的组织管理与决策方式随着信息科技与管理解决方案的发展而演进。因此，Lotus

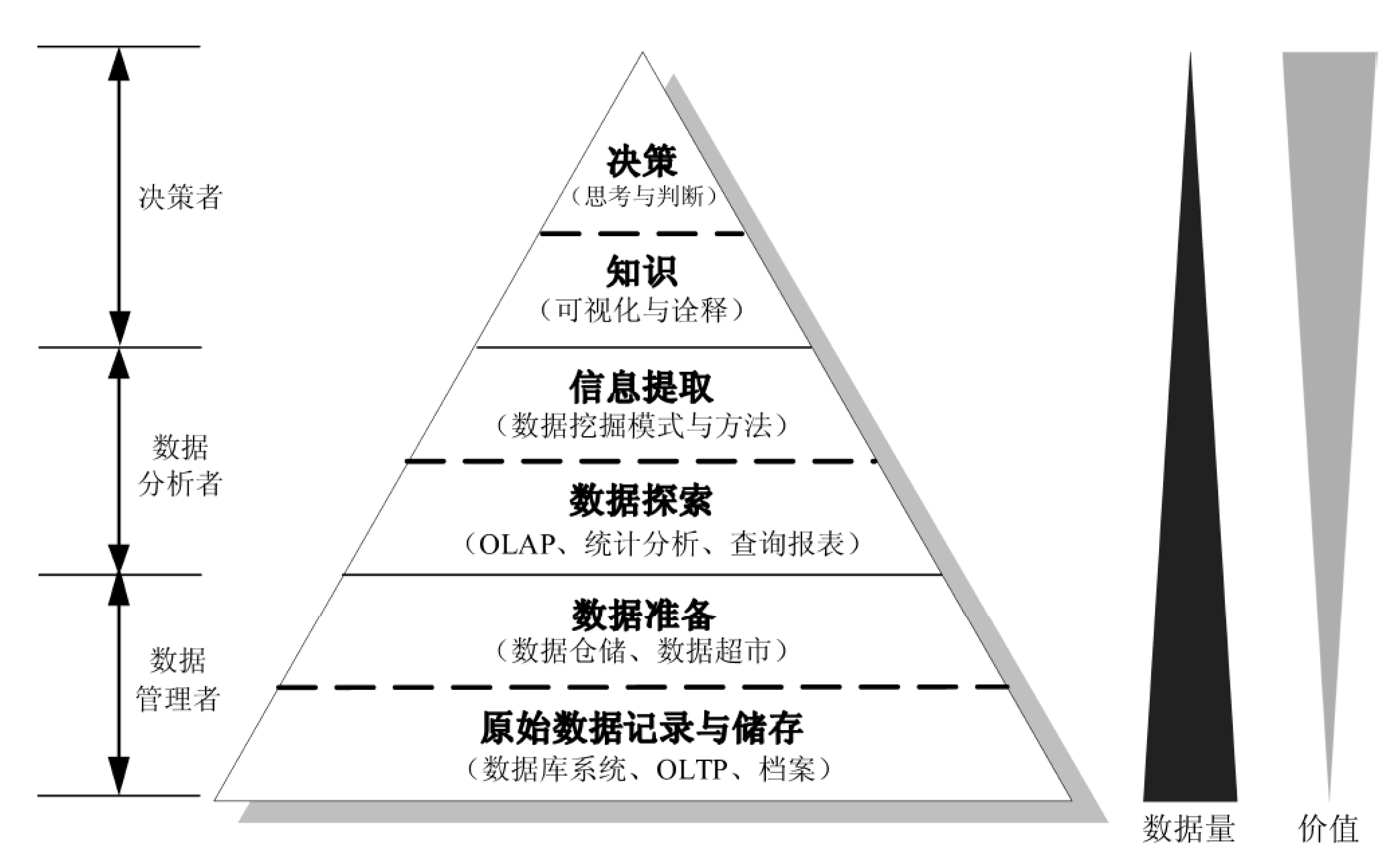


图 1.4　企业中数据的阶层分级(Cabena *et al*.,1997)

总裁帕伯斯(Papows,1999)以 16 种定位来表示企业信息与资源整合的演化过程,表 1.2 说明了数据、信息和知识对组织管理与决策的关系,其中一个维度是企业从利用数据、信息、知识到管理与决策等不同应用的层次,另一个维度是企业从强化个人、工作组、企业到供应链管理等不同范围的层次。随着其范围与应用复杂度的升高,所需要的决策信息系统就越趋复杂,然而其可创造的价值也逐渐增加,在竞争激烈的时代,企业决策信息系统演进的速度如果比对手慢,处理数据能力小,就好像用落后的武器和别人打仗,往往未战先败。

表 1.2　企业决策信息系统发展的演化过程(简祯富,2014b;Papows,1999)

	数　据	信　息	知　识	管理与决策
企业向外延伸与供应链管理的层次	供应链管理系统与应用软件	跨公司的沟通与协同	供应链的生态与公司定位	全方位的策略管理与决策
企业内部组织整合的层次	企业电子化系统与应用软件	全企业的沟通和企业整合	全企业的知识管理	企业流程与组织再造
强化工作组的层次	特殊功能软件与数据库系统	信息整理与工作组的沟通	工作组合作与知识分享	流程整合与群体决策
强化个人的层次	数据的创造、存取与使用	数据挖掘与信息提取	教育训练与知识累积	流程标准化与专业提升

1.4　数据挖掘和大数据分析架构与步骤

数据挖掘和大数据分析架构包含“问题定义与架构”(problem definition and structuring)、“数据准备”(data preparation)、“建立数据挖掘模式”(model construction)以及“结果解释与评估”(result evaluation and interpretation)四大阶段。从大数据中以自动

或半自动的方式来探索和分析数据以发掘出潜在有用的信息，此为一连串探索和重复的过程，过程中任一步骤，都可能回溯到上一步骤，不断地循环修正。首先在问题定义与架构阶段，根据问题的架构及其所做的假设(assumption)，决定数据准备的内容及格式，在数据准备阶段先行了解并归纳(induction)数据特性；然后，再由模式对数据演绎(deduction)的过程中重新整理数据的内容及格式；接下来利用建好的挖掘模式推论(inference)出影响事件变因的信息，最后再与领域专家沟通讨论挖掘结果，并检验挖掘模式的效度。如此周而复始地重复此循环将可提升数据挖掘的成果质量，并整理出可系统化的规则与模式，如图 1.5。

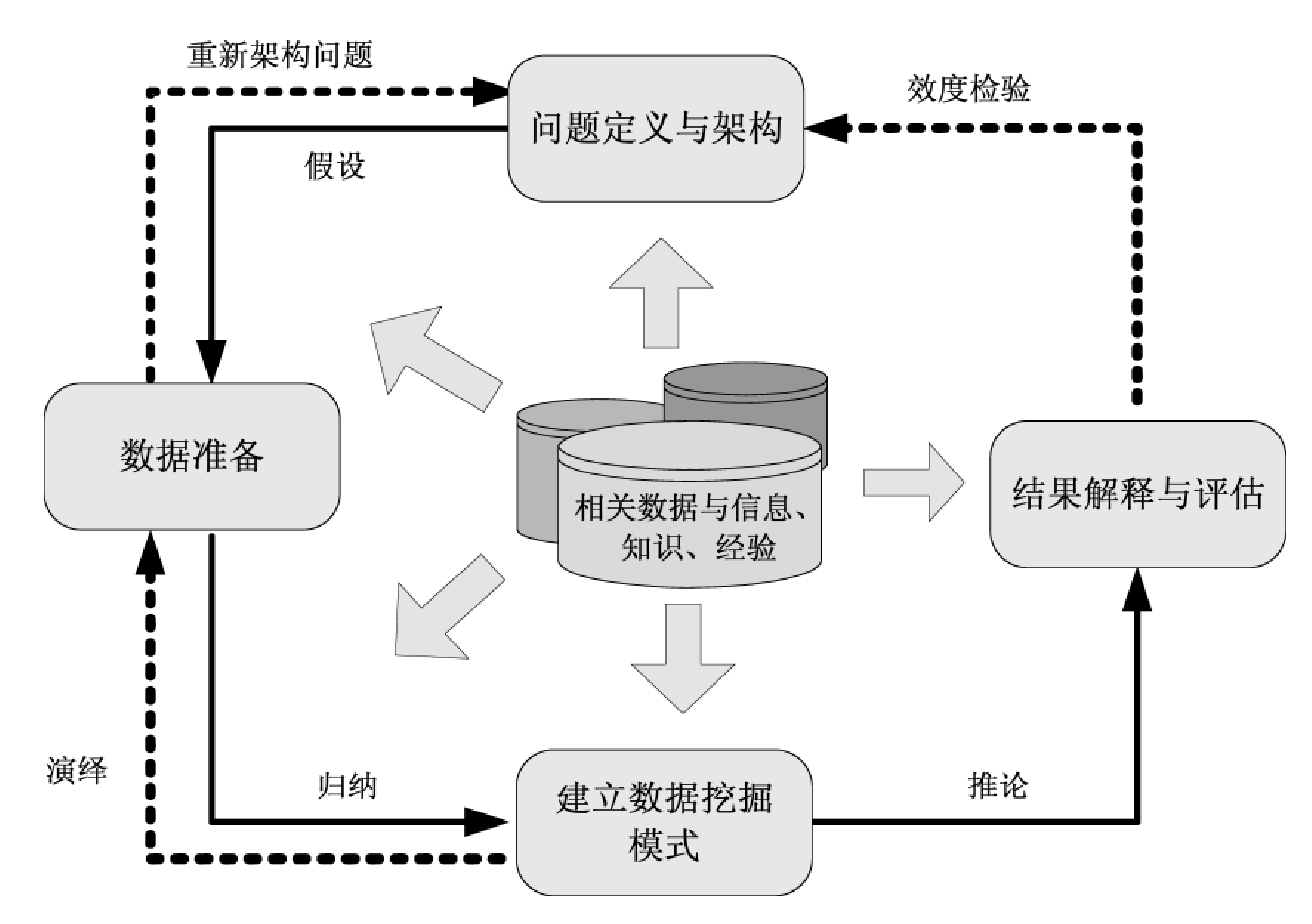


图 1.5　数据挖掘和大数据分析架构

每个阶段根据问题、数据、使用方法的不同均会影响数据挖掘的分析结果，而每一次执行后的结果也提供持续改善的循环，各阶段分述如下。

1.4.1　问题定义与架构

企业运用数据挖掘分析数据，主要是希望用以了解或找到有用的信息，或分析现况的差异，提供足够的知识以预测未来可能发生的变化。数据挖掘分析过程中须考虑数据的时间性、整合性、完整性，而不是漫无目的、“盲人摸象”似地进行数据捞取(data dredging)。为了提升挖掘效率及找到正确的挖掘方向，问题定义的阶段必须先了解问题相关的背景知识及问题特性，以清楚地陈述数据挖掘的目标，并定义试图解决的问题，将目标设定在有兴趣的挖掘对象上，如产品状况的监控(monitor)、晶圆图(wafer bin map)的分类或是低良率(low yield)产品的分析等。再依据问题定义与专业知识，选用适当的数据挖掘工具及相关分析技巧进行挖掘。数据挖掘不一定需预先设定问题的模式，所得到的结果也往往是我们先前未知的。即使遇到不同的问题类型，仍然可依照本章所提出的挖掘架构，按部就班地进行分析。累积足够的数据挖掘经验后，可以整理出系统化的规则和模式，以自动化方式进行例行分析以过滤可能发生的问题，一旦发生特殊状况，系统即能立即呈现信息，而达到系统化的最终目的。

以半导体制程事故诊断为例，半导体厂的目标主要为监控产品状况及缩短事故诊断的时间范围，以提升产品良率。然而因制造程序复杂、影响变量众多，工程师往往无法从庞大数据中迅速有效地察觉制程异常的原因，更遑论从数据中发现先前隐藏未知的重要信息。因此必须先了解半导体领域的相关知识，再根据问题的目的，搜集或回溯（retrieve）相关的制程数据，选择适当的方法或模式进行挖掘，以找出事故发生的可能原因。

1.4.2　数据准备

数据挖掘并不是将所有的数据全部盲目地放进模式中进行分析，大量数据虽可以增加发现样型的机会，但相对地，也会产生更多无用信息或噪声，影响数据处理的效能与结果的正确程度。因此，在取得数据后必须先作一连串基本的数据准备，再进行后续分析。

数据的选择通常因问题的定义而有所不同，包括判断内部及外部的可用信息，并选择哪些数据需做进一步的分析。因此，在确认问题且取得数据后，应先将数据去芜存菁，或将数据简化成分析目标时适用的格式，以确保分析数据的质量和分析结果的正确性。数据准备的目的是进一步了解数据，并过滤不当数据以确认数据格式与特性。数据的问题可分为质与量两类：质的数据可进一步细分为空间性与时间性，而量的数据则可分为训练组与测试组。在具有数据特性的概念后，即可选取欲分析的数据，进行数据转换（Pyle，1999）。至于数据准备的形式和条件，则依分析模式与所搜集的数据源不同而有所差别。

数据预处理技术主要包含，数据清理（data cleaning）、数据整合（data integration）、数据转换（data transformation）、数据归约（data reduction），详细的数据准备方式可参阅第 2 章。

（1）**数据清理**：包含遗漏值的处理、平滑（smoothing）杂乱数据、找出离群值，并纠正数据的不一致性。

（2）**数据整合**：将多个数据源中的数据结合存放在一致的数据库中。不同来源的数据可能因属性（attribute）定义或单位定义的差异，而使相同数据被误以为是不同数据，因此，必须重新检查，将相同数据放在一起。另外，也可以使用相关分析检测出冗余（redundancy）的属性，避免重复。

（3）**数据转换**：将数据转化成适合挖掘的形式。例如，分类属性“街道”时，可以将其一般化（generalization）成“地区”或“城市”。另一种方式是标准化（standardization），将属性数据按比例缩放，把原有数据置入一个小的特定区间。例如利用数据归一化（normalization）将数据转换至[0，1]区间。

（4）**数据归约**：数据的维度会影响挖掘模型的建立，一般而言，高维度的数据计算较复杂，花费的时间也较多，因此分析人员必须判断是否要进行数据归约，以降低数据维度，但同时应尽可能地保留数据的完整性，以权衡信息的保存与处理效率。

1.4.3　建立挖掘模式

选择适合的数据挖掘工具包括传统的统计分析，以及人工神经网络（artificial neural networks）、决策树（decision trees）、关联规则（association rules）、聚类分析（cluster analysis）等。例如，通过人工神经网络学习，建立制程参数数据与良率的预测模式，以预测未来制程良率；或利用决策树分析找出造成低良率的制程机台参数规则；或利用聚类分析方法，对数据进行叙述性分析，或者利用关联规则进行关联性探索。另外，根据不同数据挖掘

模式也需对参数进行设定,设定的方式可能与问题有关,例如 K 平均法(K-means)中的聚类个数 k 可能与预期的聚类数目有关;也可能需通过实验的方式来决定较佳的参数组合,例如人工神经网络中的神经元个数与网络架构。

各种数据挖掘模式的使用过程和结果应用各有不同的特性和要求,除了与决策信息系统相同的基本要求如正确性、稳定性、弹性和容易使用性外,针对处理数据的规模和速度,以及对数据的复杂性、偏差和稀疏程度的容忍能力,还有结果的再现性和可解释能力,以及内建于商业智能与决策信息系统的整合能力等,会展现出不同的数据挖掘模式特性。例如,可解释能力是指该工具得到的结果对用户而言,是否容易解读和理解。就决策树而言,决策树的结果为一树状结构,每一条由起点开始到终点的分支串联起来就是一条"若……,则……"的规则,由于其结果的可视化与规则解读的便利性,因此具有较佳的解释能力。反之,人工神经网络算法的计算方法如同人类的大脑运作般复杂,无法由结果回溯其分析过程而了解结果产生的来龙去脉,用户只能根据其结果自行判读是否具有实质意义,因此可解释能力较低。

挖掘工具端因解决的问题类型而异,每一种工具适合处理的数据类型也不相同。因此,通常需混合(hybrid)不同的数据挖掘技术以解决问题。例如,在解决企业问题时,公司可先利用聚类分析将顾客分为重要客户与一般客户等不同层级,再利用决策树分析找到不同层级客户的消费行为,作为后续目标营销的参考。

借由一开始的问题定义,可以了解大概有哪些类型的数据挖掘工具值得纳入考虑,挖掘工具本身各有所长,并没有所谓绝对最佳的方法,工具的选择与问题本身和所搜集的数据类型息息相关,领域专家的配合有时也提供数据挖掘的方法选择与改善之因素。因此,数据挖掘者本身对于工具必须具备清楚的认知以选定合适的工具。

1.4.4 结果解释与评估

针对不同的数据挖掘模式得出的结果所采用的评估指标也不同,例如分类正确性、模型误差大小、群体间的相似程度、分析所需时间等。一般来说,分析人员会评估该模式的解释能力如何、是否可接受,若不足则可能改善的方向为何,甚至可能需重新检查所搜集的数据或采用不同的数据准备方法。数据的价值在于有没有意义,并非所有分析而得的结果均有价值,在分析过程乃至于最后挖掘的结果,不论是数据、可视化图形或者规则化叙述,应不断与领域专家讨论,以获取其经验及真知灼见。人类擅长借由图像和直觉来提取有意义的信息,而可视化是最强而有力的描述方式,要找出具意义的可视化图像并不容易,但一张适当的图表,可能比几百条规则或几万笔数据更有价值。

挖掘的结果对于企业运用是否有帮助,以及整个挖掘的过程是否达到预期效果,皆须通过不断地结果解释与讨论,以厘清样型特征所代表的意义与价值,才可使研究模式与结果更加完备,之后可进一步将相同属性的规则类型储存至规则库,结合领域专家的经验与定性说明,以建立决策支持机制与知识管理系统。

总体而言,欲从庞大数据中挖得有意义的知识,除了有效的模式与工具外,事前对问题的了解、数据的准备以及事后对结果的诠释与应用同等重要。数据挖掘的结果好坏取决于对问题领域与研究目标有清楚的认知,确认具有价值的知识以及应用的目标后,建立目标数据集,再选择一个适合分析的数据集或是相关变量的子集。数据挖掘需针对问题特性与数

据类别，选择合适的数据挖掘工具分析庞大的数据，以挖掘有意义的规则或样型并整理成有用的信息；不该以使用工具为目的，强制将某工具用于不适合的问题，更不能盲目地结合数种工具并认为可以发挥加乘的效果。利用挖掘工具挖掘出结果后，需与领域专家合作以阐释挖得的信息，将所得信息以可以被确认、观察和再使用的形式呈现，使决策者能够理解，并根据所得信息回归决策的目标，拟定适当的行动方案，做出决策。最后，评估此次挖掘的成效，有效地运用挖掘结果与经验反复修正模式，改善下一个循环，并建立决策支持的机制。数据挖掘与决策支持系统的关系在于，决策支持系统是基于系统中的推论模型或经验规则提供决策上的建议与辅助，这些模型或规则可能来自于领域专家的经验或是由数据挖掘分析大量数据后，归纳而得的隐藏在专家经验后的规则或样型，而基于数据挖掘所获得的样型往往能找到原本领域专家未知的信息。

1.5 数据挖掘的问题类型

一般而言，数据分析目的可分为**描述性(descriptive)**与**预测性(predictive)**。描述性目的是希望以更易了解的方式来描述一个隐藏在大量数据背后复杂的现象或状态，借由分析数据之间的关联，找到可能的相关(correlation)、趋势(trend)、样型或规则，例如根据销售交易记录找出产品间的关联以决定促销的产品组合；预测性目的是基于历史数据的关联或规律建立模型，作为预测或判别未来的结果，例如，预估产品未来一季的销售量、判断某信用卡客户是否会有违约风险等。

数据挖掘所处理的问题类型虽不尽相同，但大致可区分为四种：分类、预测、聚类以及关联规则。

1.5.1 分类

分类(classification)是通过观察大量数据后得出规则以建立类别(class)模式，将数据中各属性分门别类地加以定义。例如，鸢尾花分类问题，利用输入花瓣及花萼的长度、宽度，通过数据分析建立区分三种不同花种的规则或模型；或者在半导体制造的良率分析中，寻找良率与制造过程中数据的关系，以制造过程的记录(使用的机台型号、通过机台的时间、在线量测参数的表现等)建立高良率与低良率的分类法则，作为判断良率好坏或诊断故障原因的方法(简祯富等，2003)。此外，图样识别(pattern recognition)也是一种分类问题，基于输入图样的输入特征，将其归类至对应的类别，例如晶圆图分类(简祯富等，2002)。贝里和利诺夫(Berry & Linoff，2004)将此类型细分为"分类"与"**估计**"(**estimation**)，其实两者意义相同。差别在于前者分类的结果属于离散(discrete)形态，后者则属于连续(continuous)形态。

1.5.2 预测

预测(prediction)是利用历史数据来预测未来可能发生的行为或现象。例如，半导体产品制程周期时间长，因此可以分析制程搜集的数据以预测产品良率，以作为优化投料量与派工决策之依据(简祯富等，2003)。预测与分类相当类似，但其中最大的不同在于其所拥有的不完整信息而造成不确定性。换言之，在预测工作中，会根据某些未来行为的预测而分类，或者估计某变量未来可能的值。要检查预测结果的正确性，只能待其发生后再加以观察与

验证。例如,Google 利用关键词检索预测流感,其结果比美国疾病管制中心的数据还快且实时。

1.5.3 聚类

聚类(clustering)是根据相似度(similarity)将数据区分为不同聚类,使同一聚类内的个体距离较近或变异较小,不同聚类间的个体距离较远或变异较大。其中,相似度可以利用不同的距离或相关(correlation)来定义。例如,依据良率高低将晶圆区分为高良率与低良率的晶圆,以辨识制程良率的状况。亦有文献定义聚类是将许多不同的群组,分成一些更相似的群组或聚类。例如,通过聚类分析了解信用卡顾客的特殊消费样型或者市场细分。

聚类与分类最大的不同在于聚类并没有预先定义好类别,聚类结果的意义须依靠分析者事后的阐释。因此,找出聚类本身,加以了解并解释聚类的意义才是最重要的工作。而聚类过程中依选择的变量不同,所得的聚类结果也不尽相同。聚类通常是在进行其他类型数据挖掘前的预先处理动作。例如,通过半导体晶圆图聚类分析,找出具有特殊样型的聚类,并针对该聚类回溯制程中造成晶圆图特殊样型的原因,以尽快排除事故原因并提升良率(Hsu & Chien,2007;简祯富等,2002)。

异常值分析是聚类分析应用的一个特性,通过相似度比对,找出与大多数聚类差异较大的样本数据。异常值的笔数或个数通常远低于其他数据,在大多数的分析情况会将异常值视为噪声而予以剔除,但当少数数据才是重要关键时,例如黄金客户鉴别、诈欺监测,异常值分析则转而成为分析重点。

1.5.4 关联规则

关联规则分析通过数据寻找分析在同一时间发生的事件(event)或记录(record),并呈现搜索结果的规则。例如,在超市顾客的交易记录中发现:“若”顾客 A 在星期五晚上买了啤酒,“则”顾客 A 同时也会购买尿布。像这样以前所未知的“啤酒—尿布”关联规则,却可以帮助超市决策者拟定交叉销售策略以促销相关商品,或变更卖场摆设方式以方便顾客选购相关联的商品来增加销售额。此外,因为半导体产品良率易受机台影响,通过关联规则分析,可以优化机台组合作为派工依据,以提升良率(王鸿儒等,2002)。通过关联规则也可寻找数据间的共通形式,例如,若晶圆在第一金属层重工且在机台甲进行蚀刻,则晶圆失误率高(Kittler & Wang,1999)。

1.6 数据挖掘模式

大数据分析的理论基础包括从分析不同问题所需的领域知识,到数据库与数据仓储记录、预处理技术以及建立模型需要的算法与数学模型,如数据挖掘、人工智能(artificial intelligence)、机器学习(machine learning)、信息检索等模式化(modeling)方法。另外,在结果解释与应用上,如何以图形或简单的可视化方法提供分析者更清晰易懂的解释方法也是有效呈现挖掘结果的关键。

数据挖掘虽属于探索驱动,不需事先假设以求验证,但需选取合适的工具或算法。挖掘的工具依需解决的问题类型与挖掘的目的而异,且通常不会只使用单一工具来进行挖掘工

作，不同的方法均有其优点与缺点，方法的适用程度与否也取决于数据的形态与种类、数据与模型应用的假设、数据集合的大小、数据噪声与数据质量、分析结果的应用目的与方式。各种模式详细的说明请见本书第 2 篇。

1.7 结论

数据挖掘的产生与信息科技的演进息息相关(Han *et al.*，2011)。由 20 世纪 60 年代的源文件搜集到发展成为数据库系统(database system)，至 20 世纪 70 年代至 80 年代初期进展到关系数据库(relational database)，数据开始以关系型数据表的方式储存，提供用户快速存取、搜索，以至于如在线实时事务处理(online transformation processing，OLTP)技术的发展。自 20 世纪 80 年代中期开始，数据库系统的研究开始蓬勃发展，连带着不同性质数据库等应用导向数据库技术逐渐成熟，另一方面，全球信息网络的出现也促使计算机科学与信息工业的快速发展。此外，硬件技术的急速成长也提供低廉的计算机，推动数据库进阶发展与数据仓储(data warehouse)，包括数据清理、数据整合与在线实时分析处理，OLAP 主要是由不同汇整角度提供数据间的统计信息，作为决策者之关联性参考，例如提供零售业者不同区域间不同品牌的消费金额差异，但若要进一步分析顾客消费行为，则需要更复杂的分析工具，如数据挖掘技术。现今，大量数据不仅改变企业经营模式，也刺激企业决策者开始思考如何有效运用数据挖掘分析技术，从各种数据中淬炼出黄金，以掌握企业竞争优势。未来，数据将成为最宝贵的资产。以网络从业者为例，若能从数以万计的消费数据记录中找到现今尚未有人发现的关键消费行为模式与可能的产品应用趋势，将可挖掘出许多未开发的潜在商机，取得市场先机。

管理大师彼得·德鲁克曾言，未来是“服务经济”(service economy)的时代，所有企业都将是服务业，在激烈的竞争环境下，能掌握顾客需求者即能掌握商机。在“顾客导向”的思维下，企业为了达到良好的顾客关系管理，必须有效地整合资源，了解顾客的需求，调整经营模式与研拟适当的行销策略，好好评估每一个顾客的需求与偏好，再针对每一个顾客提供个别的服务。借由信息科技与大数据分析的应用，发掘潜在顾客并增进与顾客间的互动，并由不同顾客群间交易记录等数据，来预测顾客需求，推荐符合顾客要求的商品或服务，持续地改善企业流程程序，以满足顾客并创造顾客价值，进而提升市场占有率。

大数据分析的能力已逐渐成为企业竞争力重要一部分。例如大型百货零售商 Walmart 利用事务数据库的分析找到公司的竞争利基，首先建立条形码扫描系统掌握每项产品的身份与相关数据，汇整全美各分店实时销售数据以分析顾客消费行为，例如，著名的“啤酒与尿布”案例。同样地，由 Walmart 的大数据分析得到的“飓风与草莓吐司饼干”是另一个著名的发现：每当飓风来临前夕，草莓吐司饼干(POP-Tarts)的销量就会随之暴增。根据此规则，一旦气象预报发布飓风消息，卖场就会事先多预备大量的草莓吐司饼干，并摆放于显眼处，大幅刺激销售业绩。

2011 年在美国知名的益智抢答比赛“Jeopardy”中，IBM 的超级计算机“华生”(Watson)打败了两位该节目史上最强的高手詹宁斯(Ken Jennings)以及路特(Brad Rutter)。华生是一台具有 2800 个中央处理器、16 兆的内存、每秒运算能力高达 80 兆次的超级计算机。要达成这项成就，华生得先听懂问题，了解题目的语意，再通过数百万条逻辑指令抽丝剥茧以

推理出正确答案。

为了监测都市热岛效应的演变情形，荷兰皇家气象研究院(Royal Netherlands Meteorological Institute，http://www.knmi.nl/)从数据分析中得出手机电池温度与环境温度具高度相关性，因此发展出极具成本效益的众包(crowdsourcing)方式，以智能手机用户安装的电池温度监测程序所搜集的温度数据，来实时监测与预测外在环境温度的变化。此外，跨国电信公司 T-Mobile 针对特定天气状况(例如，下雨)对手机信号基地台信号传输能力的影响，结合现有基地台的信号传输信息与气象预测功能，新增气象预测信息的商业模式。这些创新的应用使基地台以及几乎无所不在的手机成为简易的气象站，进一步将数据应用于农作物生产、电力需求规划等，亦省下建立气象站的大笔费用支出(Overeem *et al.*，2013)。

西班牙服装品牌 ZARA 同样运用数据挖掘与大数据分析技术，分析所销售的每一件商品，实时回复顾客信息给设计与生产端，找出顾客消费喜好与意见，帮助决策者找到时尚目标市场；优越的设计能力与强调少量、多样、迅速汰旧换新的经营风格，使其成为新一代的快速时尚王国。除了实体店铺，ZARA 也成立了多家网络店铺，用户在网络上的消费，包括浏览过的衣服、交易数量、交易金额与日期、浏览时间等都会被记录在交易信息系统，以快速整合和分析这些数据，找到不同产品的目标族群，并立刻执行商品设计、生产、配销等决策，以迅速修正与响应顾客的需求。

随着德国推动工业 4.0、美国提出先进制造伙伴计划(advanced manufacturing partnership，AMP)等，制造业的重要性再度获得关注。高科技产业如半导体制造业、TFT-LCD 光电产业等皆是高度电子化与自动化的产业，在半导体制程中会自动记录大量的数据，分析这些数据有利于进行制程质量诊断或生产力分析。但是由于数据维度不断扩张、影响因子众多而复杂，当制程发生问题时，工程师难以仅凭自己的专业知识和经验判断解答(简祯富等，2001)。这些难题包括多变量与非线性交互作用问题、间接或周期性问题、动态制程变化、制程之间的交互影响、新制程或方法的引入、产品的多样化等。仅凭个人经验的处理方式，不但易造成数据大量浪费，且影响到事故诊断和排除的效能(Chien *et al.*，2007)。而过去借由统计方法可以解决的问题，亦因数据量扩张，使问题变得复杂而难以仅用统计方法解决。因此若可以针对半导体制程的事故问题类型，运用数据挖掘技术快速而有效率地提供问题线索甚至是解决问题的根本原因(root cause)，提升工厂产品良率，将可强化企业或制造商本身的市场竞争力。

数据挖掘能处理大量数据且由数据中发掘出人类专家无法轻易辨认的特殊规则，对半导体事故诊断这类复杂的问题有相当良好的分析成效。数据挖掘针对不同的问题与半导体数据特性，如半导体制程监控、半导体制程故障诊断等，发展事故诊断分析模式与数据挖掘方法，从大量数据中探索、挖掘出隐藏信息并缩小问题范围，可作为工程师进一步解释事故发生原因的参考依据，以达到工厂事故诊断、制程改善与良率提升的目的；相同的数据挖掘架构亦可应用于企业运营的问题上，如探索技术员特质与绩效的关系，作为管理师招募聘任时的参考依据(Chien & Chen，2008，2007)。此外，数据挖掘在半导体制造的先进制程及设备控管(APC/AEC)中亦扮演举足轻重的角色(Chien & Hsu，2006)。

1.8　本书架构

本书共有 13 章，篇章架构如图 1.6 所示。第 1 篇为大数据分析与数据挖掘导论，说明数据挖掘的基本架构与各种模式与应用，以及数据准备与管理，由数据的类型开始，进而说明影响数据质量的问题与处理手法。第 2 篇为数据挖掘方法与实证，主要介绍几种数据挖掘常用的方法，包括：关联规则、决策树分析、人工神经网络、聚类分析、贝叶斯分类法与贝叶斯网络、粗糙集理论、预测与时间数据分析、集成学习与支持向量机。第 3 篇为数据挖掘进阶运用，分别探讨数据挖掘在商业智能、制造智能以及数字决策及商业分析与优化的应用。

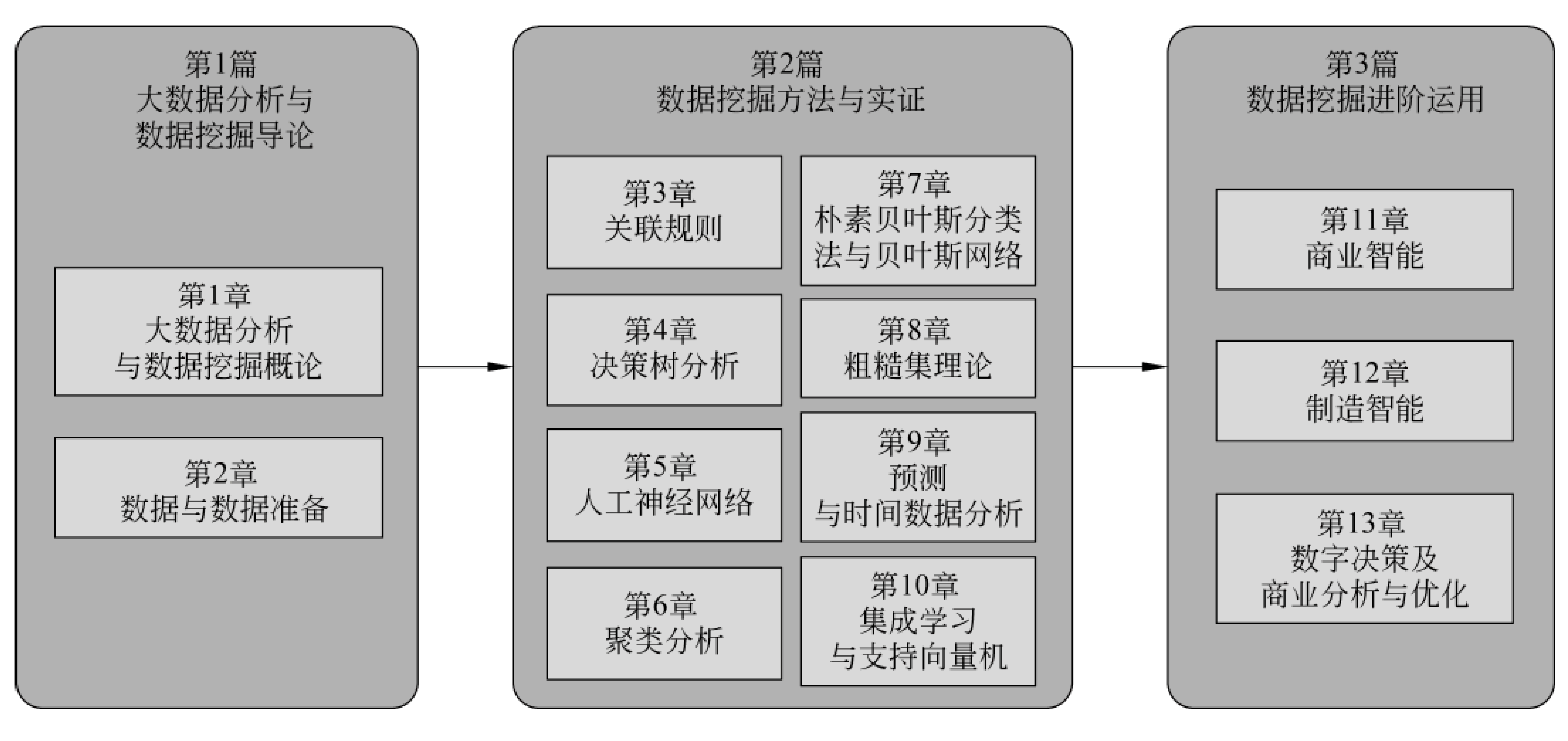


图 1.6　本书篇章架构图

问题与讨论

1. 请从网络上寻找一个应用大数据分析的实际案例，并说明大数据如何被应用。

2. 承上题，试以大数据的 4V 特性说明所寻找的实际案例。

3. 请比较统计方法与数据挖掘方法的关系，针对数据分析处理上有何不同?

4. 假设某银行推出第一年免年费的"熊猫卡"并附赠熊猫玩偶一只，发卡量因而突破 200 万张。然而，从一年后的账面数据初步分析发现其中有 15%客户领卡后从未使用，5%刷爆后列为坏账，只有 10%列为高消费无风险的"黄金顾客"（所谓的金矿）。根据上述例子，请具体详述如何利用几种特定的数据挖掘和统计分析方法由大量数据（包括顾客基本数据、每笔交易记录等）中挖掘得到可能有用的"信息"，如进一步找到重要顾客，或避免发卡给信用不好的客户。请具体详述假设可以得到的信息、需要用到的相对应工具和方法以及后续的策略。

5. 请另举一个类似上述银行业数据挖掘概念的例子，例如，电信公司的促销方案与所对应的统计方法应用。

6. 假设某银行推出年利息18%、最高可预借现金20万的“学生现金卡”，发卡量突破60万张。然而从一年后的账面数据初步分析发现其中有55%客户领卡后从未使用，15%刷爆后列为坏账，只有30%为常借钱又能持续付息还钱的金矿，请讨论银行应如何应用商业智能的方法和系统来协助管理。

7. 根据本书所介绍的各种数据挖掘工具的特性和要求，请找一个应用实例，讨论特定数据挖掘工具的关键成功因素。

8. 顾客数据是企业最大的资产，顾客数据的完备与否，对银行业务推广和获利有举足轻重的影响力。银行业多年来为了满足不同的需求而建立不少应用系统，每个系统都有其个别的顾客数据，请试着说明数据挖掘在银行业的应用，并解释可能的做法。

9. 请说明数据仓储与数据挖掘间的关系。

10. 数据挖掘的步骤有哪些？哪一个步骤比较重要，为什么？

11. 试列举三个数据挖掘在制造业的应用。

12. 试列举三个数据挖掘在零售业的应用。

第 2 章

数据与数据准备

数据质量(data quality)和数据的完整性(data integrity)决定挖掘结果的好坏。然而，由于数据搜集的方式或工具各异，导致数据库或数据仓储可能存在着许多数据噪声、数据遗漏以及数据格式不一致的状况；再加上大数据时代的数据具有数据量庞大(volume)、数据变动速度快(velocity)、数据多样性(variety)及数据真实性(veracity)等特性，若直接分析原始数据，很可能因数据质量不佳而导致事倍功半的结果或有偏误的结论。

数据准备(data preparation)是指在了解问题与目的之后，进行挖掘与建立模式之前，为确保分析数据质量和分析结果正确性所进行的数据搜集、数据预处理(data preprocessing)、数据转换及数据分割等一连串过程，以提升数据挖掘的效度和信息质量。如果数据质量不佳，如数据过度简化与无用数据太多，都会增加分析的困难度。因此，在应用数据挖掘工具进行挖掘前，需要先进行数据准备，以确保分析数据的质量和分析结果的正确性(Han *et al*.,2011；Pyle,1999)。数据挖掘工作者在数据准备的过程中，除了需与领域专家讨论及了解问题，以便选取合适的数据，也必须确保数据的质量足以进行后续分析。

数据准备的形式和条件，依分析模式不同而有所差别，一般可分成五个执行步骤：数据取得、数据检查、数据整合与清理、数据转换与归约、数据分割，以及四个维度：数据管理、挖掘效率、信息价值与工具效益。数据准备的架构如图 2.1 所示。在数据搜集阶段，必须确认

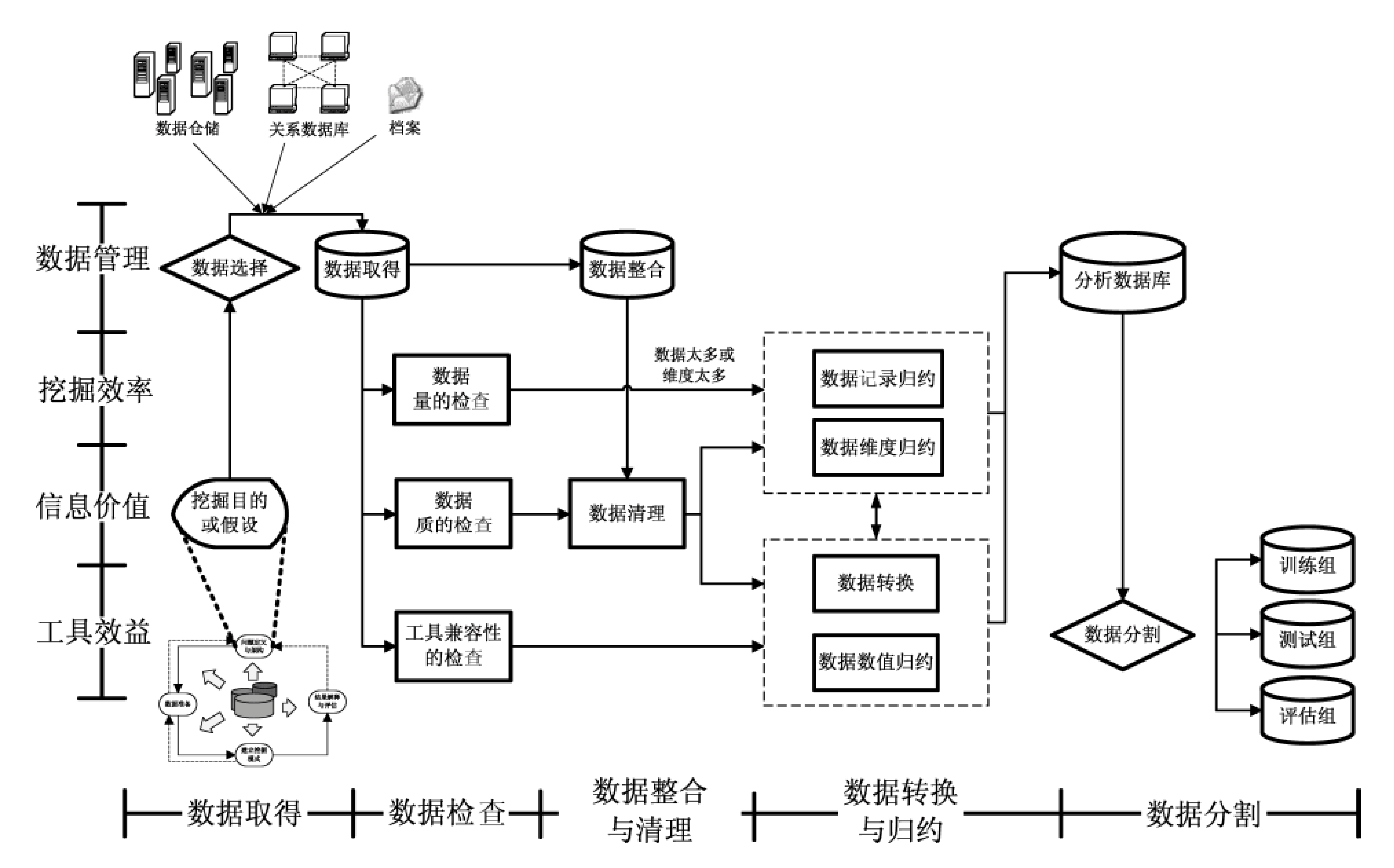


图 2.1　数据准备架构

挖掘模式所需的数据类别与取得的数据源，包含数据选择、数据获得与数据整合等步骤。数据预处理阶段则是对数据去芜存菁，删除混杂其中的不相关数据，或是将数据投射和简化以转换成适于分析目标的格式。而数据分割的目的在于建立有效、稳健的模式以及评估结果，包含数据分割与模式验证等步骤。数据准备并非一次性(one-shot)的动作，而是不断循环的过程，同时也需配合后续分析结果，直到找到合理的结果或样型为止。因此，数据准备在数据挖掘的分析过程中几乎占据了80%的时间。

2.1 数据取得

数据是数据挖掘最重要的主角。根据不同的分析目的取得数据的种类、形态也不尽相同，故需配合问题定义所得的结果进一步搜集欲分析的数据。一般而言，数据取得(data acquisition)来源可分成三种。

一、文件

文件(file)是数据挖掘的主要来源，如 Microsoft Excel、文本数据文件等，其好处是取得快速且阅读容易，缺点是一旦建立后，后续就不太容易再做数据处理，同时，若文件过多也会增加存取的难度。

二、关系数据库

关系数据库(relational databases)是由不同名称的一组关联数据表组成，每一个数据表中包含一组属性与数笔数据，也称为记录，而每一笔记录代表一个体(object)，如 Microsoft Access。关系数据库会利用个体－关系模型(entity-relationship model)来描述数据库内各属性之间的关联，并通过关系型查询语言查询数据库，例如 SQL(structured query language)即可表示两组或多组关联数据表间的关系。

零售业及大型卖场广泛使用的**事务数据库(transactional database)**即是关系数据库的一种应用，主要是记录商业交易相关的数据，每一笔记录为一笔交易结果，一般会包括交易编号、交易时间与日期、顾客编号、分店编号、消费购买物品编号等，在储存上也会利用关系数据库的架构来记录数据。

三、数据仓储

许多人容易将传统的数据库(database)和数据仓储(data warehouse)相互混淆，其实两者储存和使用数据的基本目的不尽相同。传统数据库运用数据库相关技术将过去无法处理的庞大数据都保存下来，其具有整合、保证数据质量、减少容量等优点，并以连接表格的方式读取数据，着重于单一时间的单一数据处理，为一种有系统的数据储存方式；数据仓储则储存着来自不同来源的数据，可由单一或多个数据库所组成，与数据库不同的是数据仓储中的数据大多已经过数据处理，并以“切割”的观念来读取数据，其架构如图 2.2 所示。

数据仓储利用**多维数据立方体(multidimensional data cube)**检查多维度的数据，以提供分析所需的关联分析或概念阶层的关系。多维度处理技术为事先做加总运算并把结果写入数据方块(cube)，并把方块存放在多维度在线分析处理的服务器端。图 2.3 为一个三维度的方块，右上角黑色区块“(产品，时间，地区)为(计算机，总和，北美洲)”，代表北美洲在该年度计算机相关产品的总销售量。此外，也可以视真实数据建立更高维度的数据立方体。

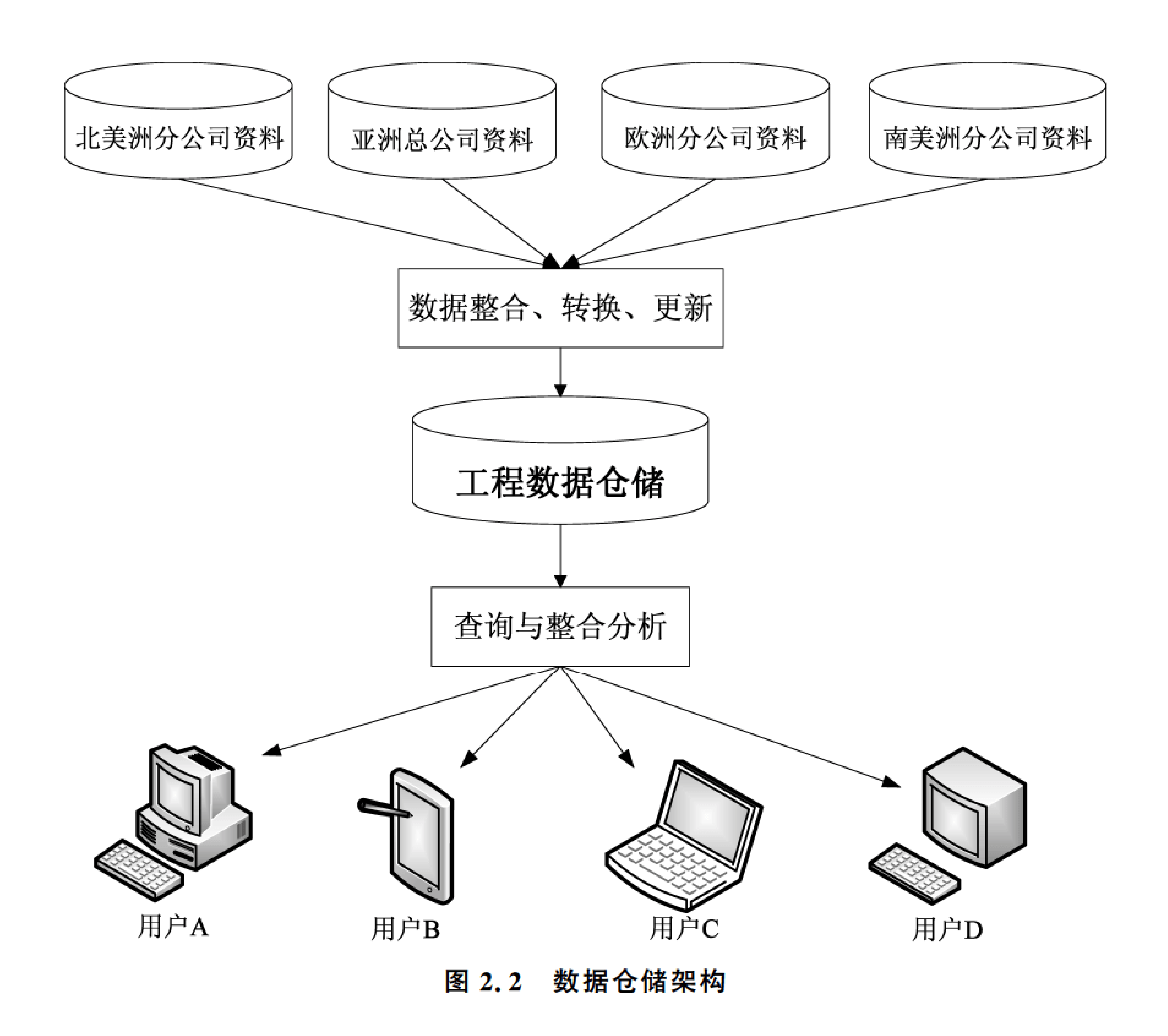


图2.2　数据仓储架构

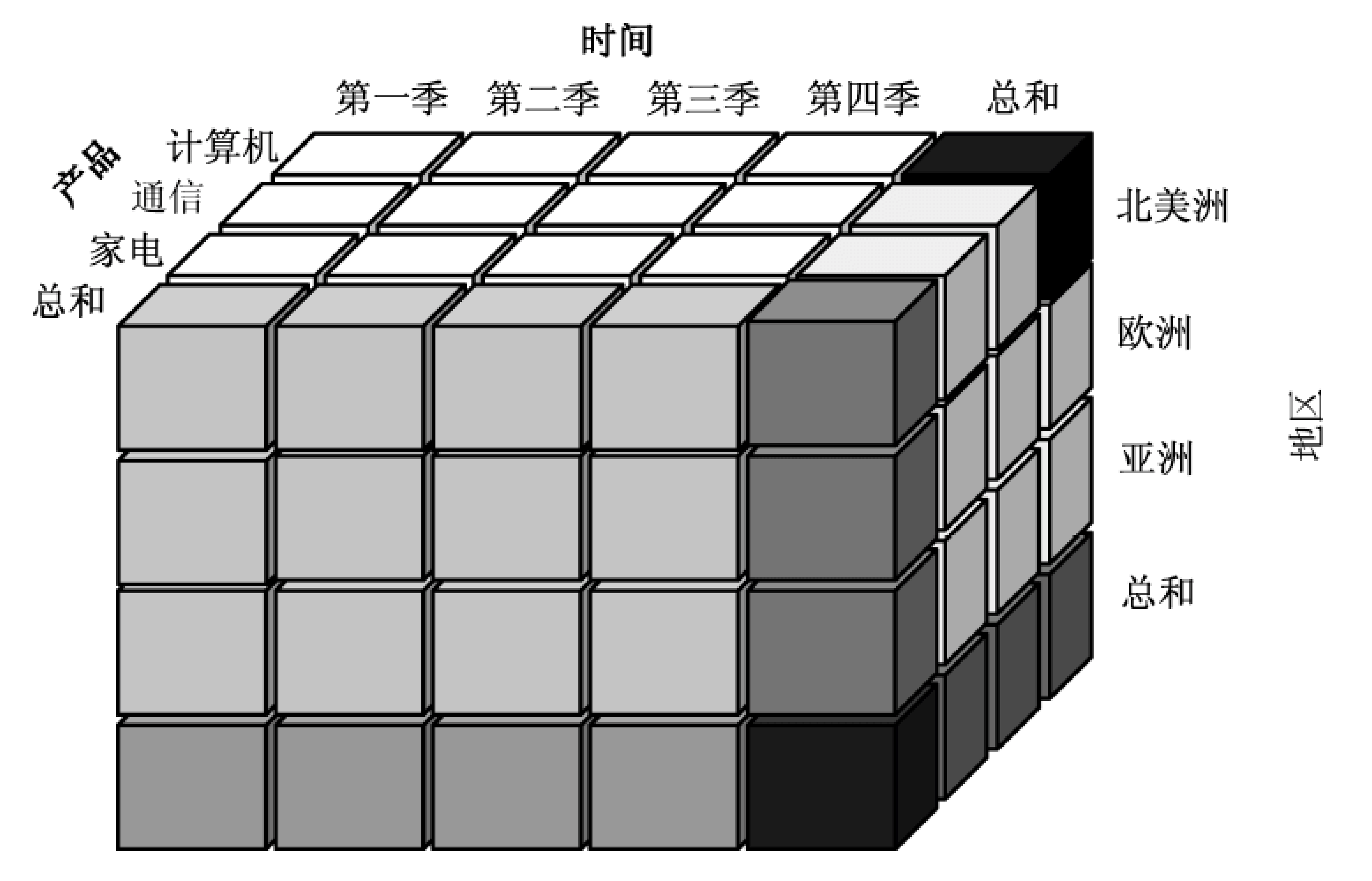


图2.3　三维数据立方体

在线分析处理(online analytical processing,OLAP)系统是一个帮助用户简易且有效率地完成多维度结构的商业数据分析工具。OLAP可筛选、分类、汇整数据仓储的数据,进而产出实体数据,再以各式数据模型呈现。OLAP包含复杂的查询功能、数据对比、数据萃取和报表,以提供不同层次的分析。用户依其专业直觉,即可通过OLAP,从不同的主题与角度操作并分析数据,得到如交叉分析、数据排名等数据,快速找出问题重点。

多维度在线分析处理系统为直接使用特殊的数据结构来执行工作,其以串行的维度作

为坐标轴,根据不同分析问题输入的条件,分析该数据库在不同构面下的关联性,提供实时查询与报表输出。在数据方块的架构下,所有数据都已事先运算并存放于方块中,快速缩减报表查询与产出的时间。但也由于必须事先算好数据方块所需的数据,考虑时间与空间资源,应避免过大的数据量。

数据仓储可以作为数据挖掘和 OLAP 等分析工具的数据源,而部分数据挖掘模式需要利用整合的、一致的和清理过的数据才能得到较好的分析结果,因此需要复杂的数据处理、数据转换和数据整合等步骤。构建数据仓储系统在进入数据存放层,也就是数据仓储本体之前,需先经过数据转换,涉及数据清理和数据整合,此构建可以被视为数据挖掘的一个重要数据预处理步骤,以避免分析工具使用错误的数据,而得到不正确的分析结果。

由于企业电子化、网络化、电子商务及云端科技的发展,企业决策者和分析师面临海量的数据。制造高度自动化的半导体厂,在制造相关数据上,已经建有完善的工程数据分析系统(engineering data analysis system,EDAS),可搜集每一段制程中的制造与质量数据;而在企业运营相关的数据上,则有企业资源计划(ERP)的各个模块储存大量数据,因此可省下不少数据搜集的时间和人力。

2.2 大数据分析的基础:Hadoop

2.2.1 Hadoop 架构

信息科技技术进步使得数据随手可得,但也造成存取上越来越困难。Hadoop 是由 Apache 软件基金会(Apache Software Foundation)以 Java 程序语言所开发的开放原始码(open source)分布式计算(distributed computing)技术,提供大数据储存与分析重要的解决方案与系统,包含分散处理环境与软件框架,以快速处理关系数据库无法处理的大数据。分布式计算的概念就是将一个工作或任务分割为多个小块,交由多台计算机共同完成一项任务,再将各台计算机的运算结果汇整而得的技术。

Hadoop 能有效处理大量的数据并具有提供储存的能力,同时可整合多台计算机的资源,提供数据分散运算,在极短时间内即可完成运算工作,并且自动保留数据副本,提高数据的可靠性与延展性。

Hadoop 架构的两个核心主要包括:①Hadoop 分布式文件系统(Hadoop distributed file system,HDFS),将数据进行切割并制作副本备份,再分散储存于不同的计算机或服务器上,提供数据的快速存取,并且有效备份在不同的硬件以避免数据损坏;②Hadoop 分布式计算处理架构(MapReduce),MapReduce 是由 Map 与 Reduce 所组成,Map 主要是将数据分散计算,Reduce 则是整合 Map 计算后的结果,提供分布式的数据平行处理分析。除了 HDFS 与 MapReduce 外,根据 Hadoop 所延伸的其他项目,已发展成为一个生态系统(ecosystem),如图 2.4,包括 Avro、Hbase、Hive、Pig、Sqoop、Zookeeper 等(http://hadoop.apache.org/),说明如下(White,2010):

- Avro:提供有效率的跨程序语言远程过程调用(remote procedure call,RPC)的数据串行化系统。
- Hbase:是以字段(column)为基础的分布式数据库系统,用以储存大量数据,提供快

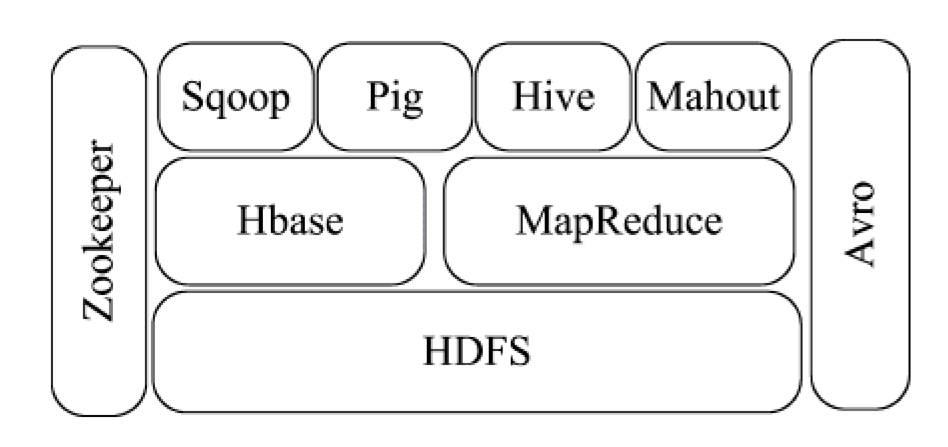


图 2.4　Hadoop 生态系统(数据源：http://hadoop.apache.org)

速的数据读取与写入。

- Hive：分布式数据仓储,提供类似 SQL 的查询语言以查询数据。
- Mahout：提供数据分析所需的机器学习(machine learning)与数据挖掘链接库。
- Pig：提供大量数据集的处理与执行。
- Sqoop：提供数据能有效率的在关系数据库与 HDFS 之间转换。
- Zookeeper：提供分布式应用处理的高效率协同服务。

2.2.2　Hadoop 分布式文件系统

HDFS 是根据 Google 文件系统(Google file system,GFS)发展而来的系统(Ghemawat *et al.*,2003)。HDFS 为采用串流数据存取模式的分布式文件系统,用以储存大型数据集,可建立在一般的硬件环境下,通过数千台硬设备的串连实现,而不需要昂贵的硬设备。即使其中有些硬设备无法运作,整个 HDFS 仍能继续正常运作(Borthakur,2008)。过去需将数据整合至同一分析数据库或数据集进行分析,在大数据的储存与分析时,数据的移动是耗时、不易且高成本的,因此,HDFS 将运算程序移动至靠近数据所在的硬设备,以节省成本与运算效能。

区块(block)是一次读取或写入的最小单位。在 HDFS 中将文件切割为相同大小的区块,一般为 64 MB 或 128 MB,为了避免区块、磁盘、设备故障,区块都会备份至其他硬设备上,如果区块检测到错误而无法使用,会由其他硬设备上读取另一个备份并执行数据回复,而在 HDFS 中区块的文件储存预设为 3 份,此设定可由程序开发人员修改。

HDFS 通常包括许多丛集(cluster),而一个 HDFS 丛集是由 namenode 与 datanode 以 master-worker 的模式运作而成(White,2010)。namenode(master)负责管理文件系统的 namespace,以维护其文件和目录的 metadata,datanode(worker)是负责储存数据,用户只要通过 namenode 即可知道文件被分割为哪些区块以及区块被划分至哪些 datanode。

假设有一客户 A(client A)想要将一笔大量数据集储存至 HDFS 中,该数据集被划分为 A、B、C、D 四个区块,而 HDFS 丛集中包括一个 namenode 与 8 个 datanode,其中 datanode 两两分布于 4 台硬件机架(rack)中,如图 2.5。在写入文件前,客户会先询问 namenode 可将 A、B、C、D 写入至哪几个 datanode,根据 HDFS 区块复制的规则,其中两份在相同机架上,另一份在不同机架上,以避免机架的毁损。以区块 A 为例,假设可写入 datenode1、datenode2、datenode5,总共三份区块分别储存在机架 1 与机架 4 中,接下来依次完成其他区块(B、C、D)的写入。假设现有客户 B(client B)想要读取该笔大量数据集,首先向 namenode 查询 A、B、C、D 所在的 datanode 信息,回复结果:"区块 A 位于 datenode1、datenode2、datenode5,区块 B 位于 datenode3、datenode4、datenode5,区块 C 位于

datenode1、datenode2、datenode6，区块 D 位于 datenode3、datenode7、datenode8"，考虑离客户最近的 datanode，namenode 依序从 datanode5 读取区块 A、从 datanode5 读取区块 B、从 datanode6 读取区块 C、从 datanode7 读取区块 D。当区块 A、B、C、D 都读取完毕后，即完成该大量数据集文件的读取。

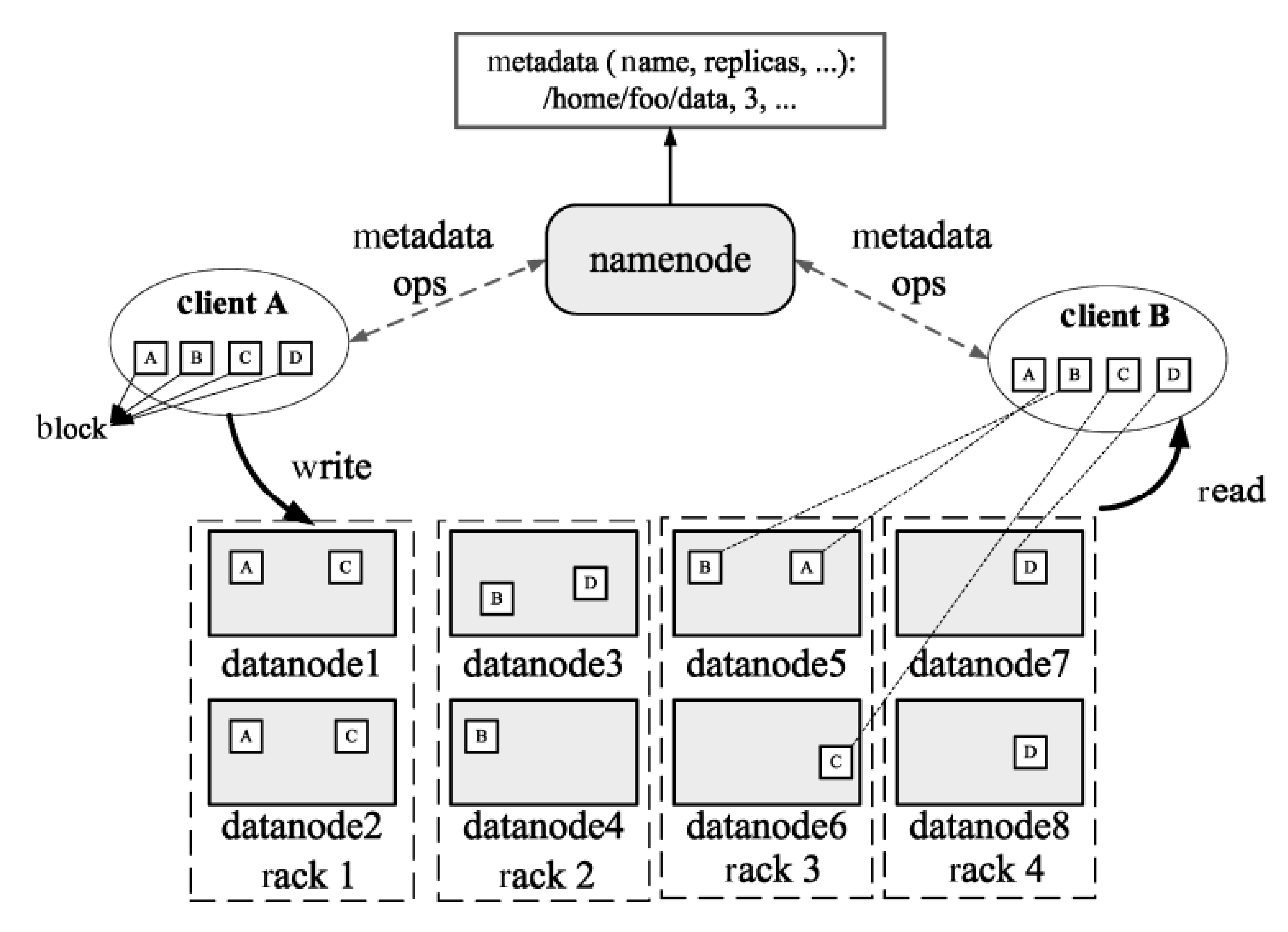


图 2.5　HDFS 架构(图片来源：修改自 Borthakur，2008)

2.2.3　MapReduce

MapReduce 是一个分布式的程序架构(Dean & Ghemawat，2008；Dean & Ghemawat，2004)，采用分治法(divide and conquer)的概念，将运算任务分割为许多小的任务后个别处理，之后再做加总。分割的目的在于利用多个机器运算以获得较好的负载平衡，同时所花的时间也远低于一次处理全部数据的时间，但分割的容量太小则会造成文件管理与建立 Map 任务的负担，一般而言，分割的大小应该与 HDFS 的区块大小一致(White，2010)。

MapReduce 将处理程序(process)分为 Map 和 Reduce 两个阶段，每个阶段的输入与输出都采用序对(key，value)，程序开发人员需要撰写 Map 函数与 Reduce 函数，作为大量数据集运算任务的平行处理。如图 2.6，首先将输入数据划分为多个小分割(Split)，处理的任务也分为多个子任务，在 Map 阶段则将待执行的子任务与分割合并处理，经过排序、复制、合并后并产生中间数据，在此过程又称为洗牌(shuffle)，在 Reduce 阶段则将所产生的中间值数据汇整为最终结果。

假设有一组数据为{牛奶、面包、柳橙汁}、{面包、饼干、饼干}、{柳橙汁、饼干}，共分为三个分割，在 Map 阶段定义输入与输出为(项目：个数)，以第一个分割为例，根据其输入项目(牛奶、面包、柳橙汁)，可得到输出为(牛奶：1)、(面包：1)、(柳橙汁：1)，其余两个分割也经过 Map 函数转换，最后将所得到的中间数据经由排序、复制、合并后产生(牛奶：1)、(面包：2)、(柳橙汁：2)、(饼干：3)，并将其作为 Reduce 函数的输入，最终得到结果为(牛奶：1；面包：2；柳橙汁：2；饼干：3)，如图 2.7 所示。

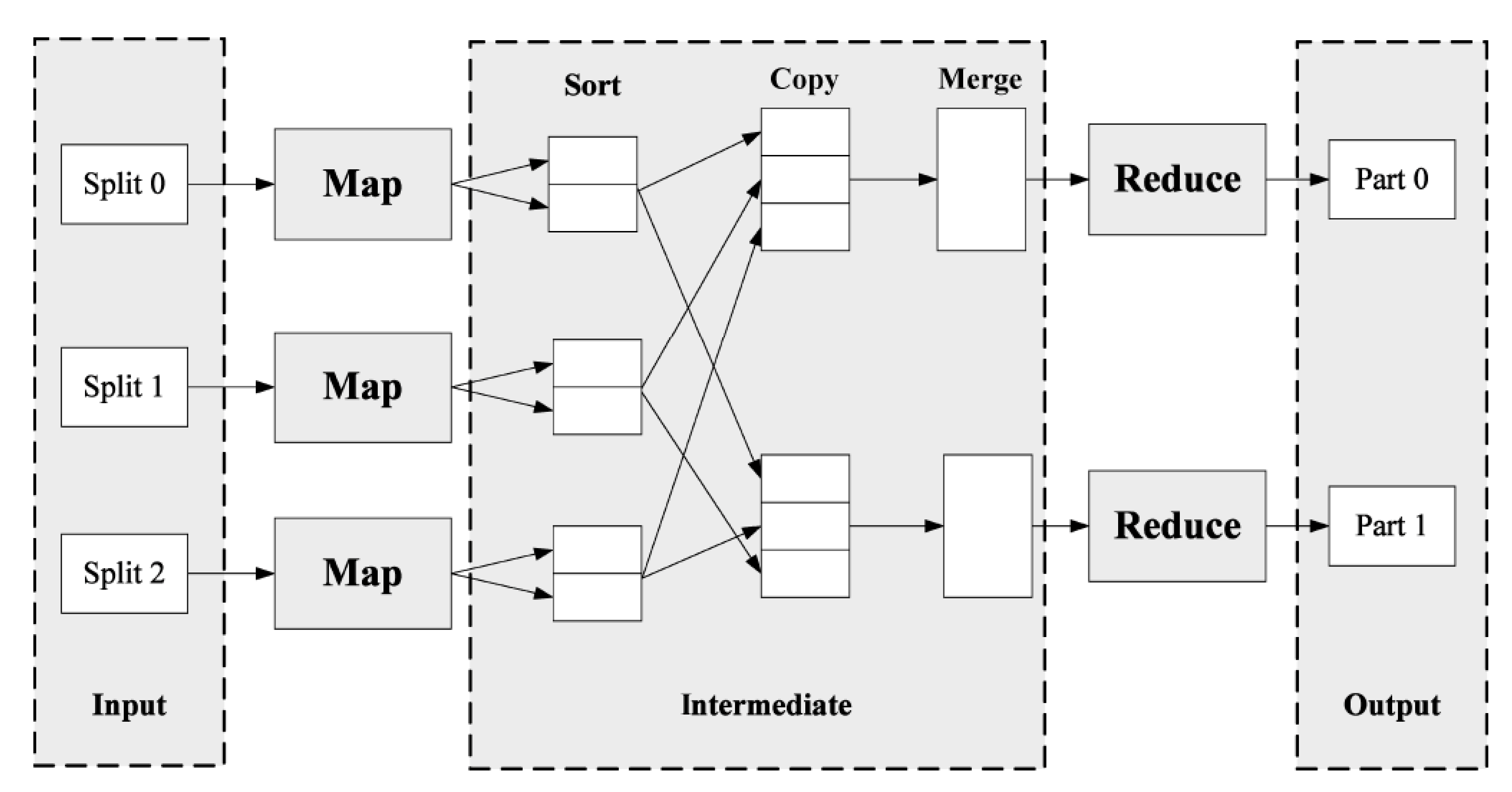


图 2.6 MapReduce 架构(数据源:修改自 White,2010)

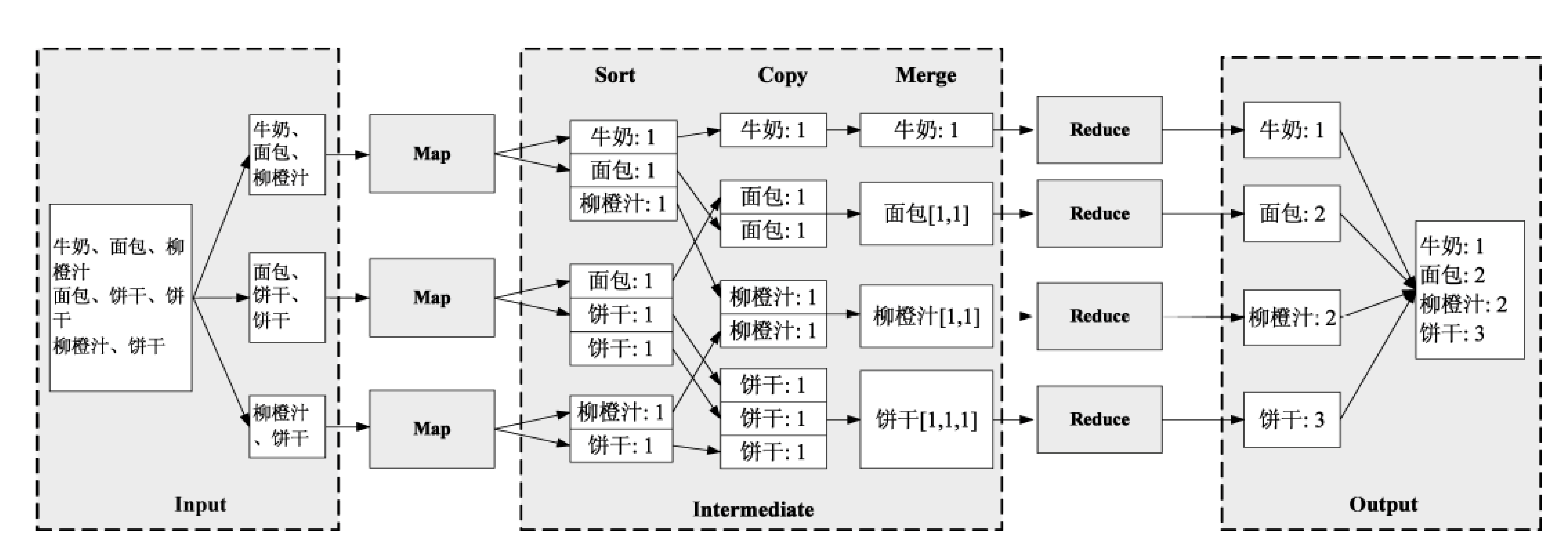


图 2.7 MapReduce 范例

2.3 数据类型

数据可以是一笔数值图形、向量、文字或案例等记录;而数据的构成与形态则包括数值、字符串、布尔值以及日期等。其定义与特性分述如下:

(1) **数值**:数值数据为最常用的一种数据类型,数据储存内容为数值形态,例如整数与实数,可用来储存如年龄、良率、员工年资、货品销售量等数据。

(2) **字符与字符串**:字符串是字符的数组;字符串数据形态即是储存一串互为相同或不同的字符,例如以“男”与“女”字符记录性别;以“张三”与“李四”字符串记录姓名。

(3) **布尔值**:布尔数据只有两种值,分别为真(true)与伪(false),通常用来储存一些可供程序判断的条件结果,例如以“真”与“伪”分别记录满 18 岁的人与未满 18 岁的人。

(4) **时间性数据**:时间性数据是数据本身或数据库中含有时间前后或顺序相关的特性,专门用于提供日期等相关操作,储存方式包含年、月、日,或更精准的时、分、秒等,为某一时间下的关联数据记录,借由时间来标记该笔或该组数据的发生时间。例如花店的销售记

录说明在某个时间点每一位顾客购买的花束,整合分析后发现情人节前夕是玫瑰花销售的高峰。若进一步列出每位顾客在不同时间点购买的花束,将可能发现"购买玫瑰花的人,往往也会购买康乃馨"的样型。此外,有些数据本身还有先后顺序关系则称为**序列数据**(**sequence data**),例如基因序列虽然没有时间标记,但数据本身即由固定的顺序所组成,或是网站的点选顺序数据,各网站彼此间可能为依序发生的关系。时间性数据中最典型的是**时间序列数据**(**time series data**),记录着一段时间区间的结果,其特征是每笔数据会受到时间增加而改变,也就是数据间彼此相关,例如某一只股票的每日股价、台湾地区每天的气温等。

(5) **空间数据**:为数据中包含空间(spatial)相关的属性,例如亚洲区域的气温数据,即包括不同经度与纬度下的气温,又如 Google Map、地理数据库、集成电路设计规划(integrated circuit design layout)、晶圆曝光规划(wafer exposure layout)等。空间数据随着网络科技与全球定位系统(global positioning system,GPS)技术的进展,逐渐发展出越来越多应用,例如应用地理数据库于车辆导航,在上下班高峰时间避开塞车路段,或者根据当下的位置,推荐附近餐厅与提供停车场信息等。

(6) **文本数据**:其特征为将文本(text)的段落叙述加以利用,常见的文本数据报括专利报告、诊断报告、笔记、产品规格书等。其可分为:①结构化数据,如图书馆书目编辑数据;②半结构化数据,如电子邮件、XML(extensible markup language)网页数据;③非结构化数据,如社交媒体微博上的留言。文本数据的处理称为文本挖掘,常见的应用包括文件分群、摘要撷取。此外,由于文字本身有一定的意义,在分析上也会需要字典或特定名词库来协助判读词意或语意。

(7) **多媒体数据**:包括图片、声音及视频等,相较于其他数据类型,多媒体数据(multimedia data)的文件大小一般都非常庞大,在数据的储存与搜索上均需要特殊的方法,例如数据压缩(data compression)。

2.4 数据尺度

数据的每个因子都有对应的**属性**(**attribute**)及其**衡量尺度**(**scale**),以具体量化和衡量不同数据在该因子的水平(level)。例如,减肥的目标,可用体重作为衡量属性来比较不同减肥方法的成效,体重可用千克或磅为尺度来衡量减肥目标达成的程度。当某个因子不容易找到对应的属性时,可以找到相关的**代理属性**(**proxy attributes**)作为衡量。例如某光电公司曾委托作者执行某良率提升计划,当时欲衡量发光二极管(LED)的良率,却发觉无法以"个数"作为衡量属性,因为 LED 体积虽小,但每一批(batch)的产量却相当庞大,不易逐颗盘点,且经检查为不良品的 LED 会从该批中取出放到不良品区;因而改以"重量"作为衡量属性,只要知道每一颗 LED 的重量,再将每一批 LED 良品和不良品分别称重后即可估计其良率。因此,在数据挖掘的过程中应充分了解数据的特性和管理的含义。

当被衡量的对象有一个自然形成的公认尺度即采用自然量化尺度(natural quantitative scale),例如,衡量时间可以使用分钟、小时,衡量距离可以使用千米、海里等;当被衡量的对象没有像公制或英制一样自然公认的尺度时则采用定性尺度,例如,空气质量、顾客满意度、TFT-LCD 显示器的彩度偏好等,必须依据一定程序来进行尺度构建(scale construction)以

发展一套有效的尺度，把人的主观判断萃取出来后，再用某个尺度与单位来叙述它，方案衡量所得的数据才有意义。例如，社会科学常用的李克特量表(Likert scale)可以用来衡量客户满意度。要有效地构建定性数据的衡量尺度非常困难，通常只能用名目尺度或顺序尺度来衡量，例如，民意调查时，通常针对满意或不满意等不同水平的响应做编码后，再加权计算其平均值。经过严谨过程所建立的尺度，也可以作为其他相关决策的参考。例如，医学上以巴塞尔指数(Barthel index)判断老年人的行为能力，并决定是否需要聘用看护；所以不论看诊的医生是谁，根据量表所评估的结果应有一致的可靠度。

有些评估属性可以找到可能不止一种有意义的衡量尺度，例如，衡量体重的尺度可以是千克或磅。不同的尺度之间亦可互相转换，例如，一千克等于2.2046磅。以下将逐一说明常用的六种尺度。

(1) **名目尺度(nominal scale)**：名目尺度下所衡量的数字仅是作为代码来确认方案，数字的大小不具任何意义，也不能做数学运算。例如，以学号或身份证号码代表某一个人，投标厂商的代码等。

名目尺度所衡量的数字转换时必须保持数字上的代码对应关系。例如，学号或身份证号码等名目尺度所衡量的数字不会有重复的情形，每个数字仅代表一人，而每个人也只会有一个数字代码。因此，有意义的转换方式必须是做一对一的转换：

$$x_j \in X,\quad V(x_i) \neq V(x_j) \Leftrightarrow W(x_i) \neq W(x_j),\quad \forall x_i$$

(2) **类别尺度(categorical scale)**：类别尺度是将欲评估的方案依其特征分类，再将每一个类别标识一个数字代码，所衡量的数字仅是用来表示其归属的类别，因此类别尺度的数据可以重复。例如，住址中的邮政编码、电话号码中的区域号码。

类别尺度和名目尺度一样，有意义的转换方式必须是做类别代码一对一的转换，以保持类别数字代码的对应关系如下：

$$x_j \in X,\quad V(x_i) \neq V(x_j) \Leftrightarrow W(x_i) \neq W(x_j),\quad \forall x_i$$

(3) **顺序尺度(ordinal scale)**：顺序尺度下所衡量的数字表示方案之间的大小顺序关系。例如，依据进入公司先后顺序排列的员工工号、比赛名次、产品质量的等级等。顺序尺度下第二名仅表示没有第一名好而已，而且第一名和第二名的差距，也不一定等于第五名和第六名的差距。

顺序尺度的转换必须保持其数字上的大小顺序关系，因此必须以严格递增函数的方式来作转换，例如：

$$x_j \in X,\quad V(x_i) > V(x_j) \Leftrightarrow W(x_i) > W(x_j),\quad \forall x_i$$

(4) **间距尺度(interval scale)**：间距尺度所衡量的数字可以有意义地描述并比较数字之间的差距大小，又称为距离尺度(distance scale)。例如，衡量温度的尺度，摄氏温度是将水的冰点和沸点分别定为0℃与100℃，再将中间的差距分为100等份，每一度的差距相等，所以49℃和50℃之间的温差与85℃和86℃间的温差相等。然而，间距尺度并无固定原点(origin)，可以随意调整原点位置，也可以调整分隔的间距大小。例如，华氏温度和摄氏温度的零度就不相同，而且还有比华氏温度或摄氏温度的零度更低的温度；换句话说，间距尺度的原点设定不是绝对的。

间距尺度的转换必须保持其数字之间的间距大小关系，因此有意义的转换方式必须是线性函数，例如：

$$V(x_i)-V(x_j)>V(x_k)-V(x_l)\Leftrightarrow W(x_i)-W(x_j)>W(x_k)-W(x_l),$$
$$\forall x_i,x_j,x_k,x_l\in X$$

因此,$V(\cdot)=aW(\cdot)+b$,其中 a,b 为常数。

间距尺度的数值仅可进行加减运算,因此,我们不能说 100℃是 50℃温度的两倍热,因为将温度调整为华氏尺度后,数字上 212℉就不是 122℉的两倍。事实上,间距尺度所衡量的数字之间的变化和差距,比数值大小更重要。

(5) **比率尺度(ratio scale)**:比率尺度所衡量的数字之间可以做比率倍数之间的比较。例如,拿一支笔作为标准单位,以最原始的方法一段一段量,就可以得到某一面墙的宽度相当于三十支原子笔,也就是墙和原子笔的长度之间有三十倍的比率关系。比率尺度还包括重量、货币面额、时间长短等单位。

比率尺度有固定的原点,因此不同单位的任意二个值,其比率完全相同,例如,美金 1000 元为美金 500 元的两倍,转换成人民币后仍然维持两倍的关系。比率尺度的数值可进行加减乘除运算,其兼具间距尺度的特性,因此也可以有意义地描述并比较数字之间的差距大小。比率尺度的转换必须保持其数字之间的比率大小关系,因此有意义的转换方式必须是倍数关系,例如:

$$V(x_i)/V(x_j)>V(x_k)/V(x_l)\Leftrightarrow W(x_i)/W(x_j)>W(x_k)/W(x_l),$$
$$\forall x_i,x_j,x_k,x_l\in X$$

因此,$V(\cdot)=cW(\cdot)$,其中,c 为常数。

(6) **绝对尺度(absolute scale)**:绝对尺度所衡量的数字具有绝对的意义,因此无法再做其他有意义的转换。例如,概率值。

较精细的尺度除了包含较粗略尺度的性质,也可简化为较粗略的尺度,例如,比率尺度所具有的顺序尺度性质可表达决策者的偏好顺序,而所具有的间距尺度性质可用距离来表示偏好的差异大小;反之,越粗略的尺度则不包含精细尺度的性质,更不能转换为较精细尺度。因此,若数据特性许可,应选择较精细的尺度来搜集数据,以利后续的分析应用。不同类型的衡量尺度可以允许不同的运算和结果解释,且由于分析工具或是解决问题观点的不同,因此需先了解尺度的类型,再对原始数据形态加以转化或编码,以配合所用的分析工具(Pyle,1999)。例如,测量参数形态的数据在一般情况下多为间距尺度,但是有时顾及问题定义时的分析方向,会将间距尺度转化成名目尺度,例如,根据领域专家的建议将参数值大于某个数值以上的产品视为良品,反之则视为不良品。如此的数据转换对于数据挖掘的结果好坏可能有极大的影响。

2.5 数据检查

获取的数据往往不见得可立即适用于后续的挖掘分析,因此,对数据进行前置处理将使得后续在挖掘时更容易发现有意义的结果或样型。其中,**数据检查(data inspection)**是数据预处理的第一个步骤,以找出有问题的数据,并以不同的维度来检查所获得的数据,以便能事先观测出其中的错误,并与领域专家讨论以决定是否修正或删除其中数据。

数据检查可分为数据的数量与质量两方面。数据数量方面应检查量化数据的三个维度:样本个数、属性或特征个数、不同的数据值。例如,样本个数太少会影响结果的解释程

度，若数据的搜集成本不高，可试着再次搜集数据；当个数太多时，则统计上的显著不见得有实质意义。

数据质量的检查可利用数据的集中趋势以及变异程度(dispersion degree)。集中趋势衡量方法包括了平均数(mean)、中位数(median)、众数(mode)等；当得到一组数据时，通常会希望通过几个重要的特性来描述这组数据的分布状况，如大部分的数据集中在何处，数据分离的程度与范围有多大(离散趋势)，数据的分布是不是有偏向左边或右边(偏态系数，coefficient of skewness)，数据的形态是不是在某些地方特别呈现较高的频率(峰态系数，coefficient of kurtosis)。此外，可利用叙述性测度(descriptive measure)包括位置测度、变异性测度、偏态测度与峰态测度来综合样本数据的信息所整理出来的特定数值，以描述数据中的集中趋势以及变异程度。

变异程度则可利用标准差(standard deviation)、四分位距(interquartile range，IQR)、全距(range)或是变异系数(coefficient of variation)等进行衡量，并应考虑数据的完整性，如数据分布的一致性、数据定义上的偏差、数据拼写错误等；数据遗漏(data missing)，如数值或变量数据遗缺、不一致的数据等；数据噪声(data noise)，如离群值和噪声数据等。若所分析的数据为时间序列数据时，则需检查数据的季节性(seasonal)、趋势性(trend)、循环性(cycle)等特征。针对不同的数据质量，可利用相对应的检查方法。例如，以折线图或散布图检查遗漏数据与数据趋势，以盒须图检查离群值。

2.6　数据探索与可视化

原始数据经整理后，按特定规则制成表格，以系统化的**统计表(statistical table)**表格呈现复杂的数据集，或以**统计图(statistical chart)**来表示统计数据各项特征，让数据分析人员更容易了解数据的分布情形或隐含信息。

1. 盒须图

盒须图(box-and-whisker plot)亦称箱型图(box-plot)，是利用图形显示数据的中央趋势与离散程度，如位置测度与变异量数，检验数据的极端量数及分布形态。

盒须图主要构成包括中位数(M_d)、第一四分位数(Q_1)、第三四分位数(Q_3)、最小值以及最大值，如图 2.8 所示。盒子的下界限为 Q_1，也就是数据的第 25 个百分位数，上界限则是 Q_3，也就是数据的第 75 个百分位数，因此盒子的长度 Q_3-Q_1 也就是四分位距，盒中包含有 50%属性的数据，所以盒子长度越大，代表数据的分散情况越大。由盒子上下界所延伸出的线，称为须(whisker)，用以连接离群值(outlier)与极端值(extreme)的最大值与最小值。当数据介于 1.5 倍至 3 倍 IQR 之间，称为离群值；而超过 3 倍 IQR 的数据，则称为极端值。

2. 折线图

折线图(line chart)是由一条线连接数点以显示序列，并以图表的方式呈现数据分布的变化趋势，用户可以由折线的上升或下降清楚看出序列的变动，推测数值的变化，通常用来比较一段时间的数据变化或两序列以上的变动情况。其中，纵轴代表测量值，横轴代表类别目录卷标。

图 2.9 为以表 2.1 的 A、B 两公司 2000 年至 2007 年的年利润率历史数据绘制成的折

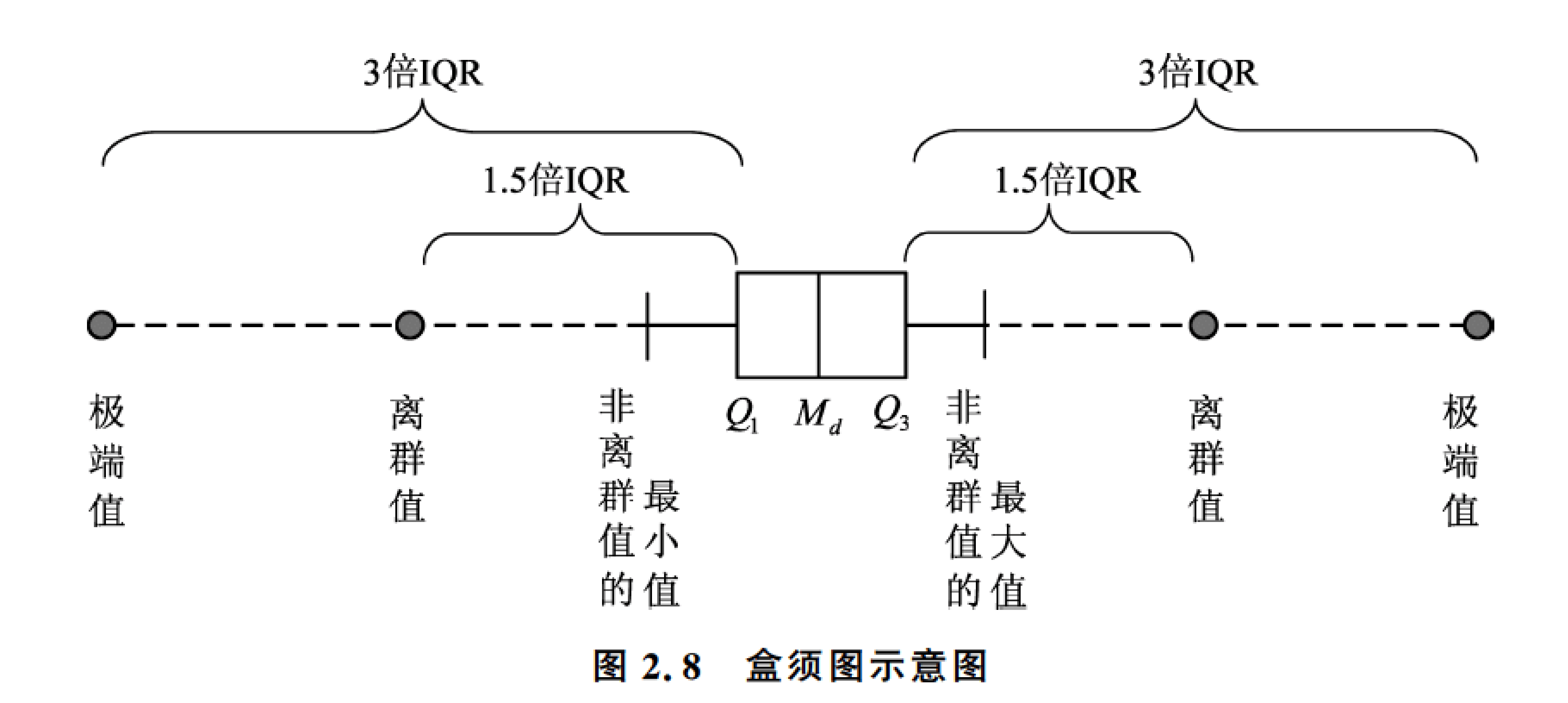


图 2.8 盒须图示意图

线图，由此两时间序列的变动情形可看出 A 公司近八年的年利润率一路下滑，表示运营状况出现问题；而 B 公司近八年的年利润率则呈现 W 形，可进一步分析其中潜藏的趋势或循环因子。

表 2.1 A 与 B 两公司的年利润率时间序列数据

公司 \ 年份	2000	2001	2002	2003	2004	2005	2006	2007
A	15.5%	12.5%	11.6%	11.2%	10.5%	9.7%	8.5%	8.0%
B	22.5%	18.9%	16.7%	12.1%	13.8%	10.6%	15.2%	17.9%

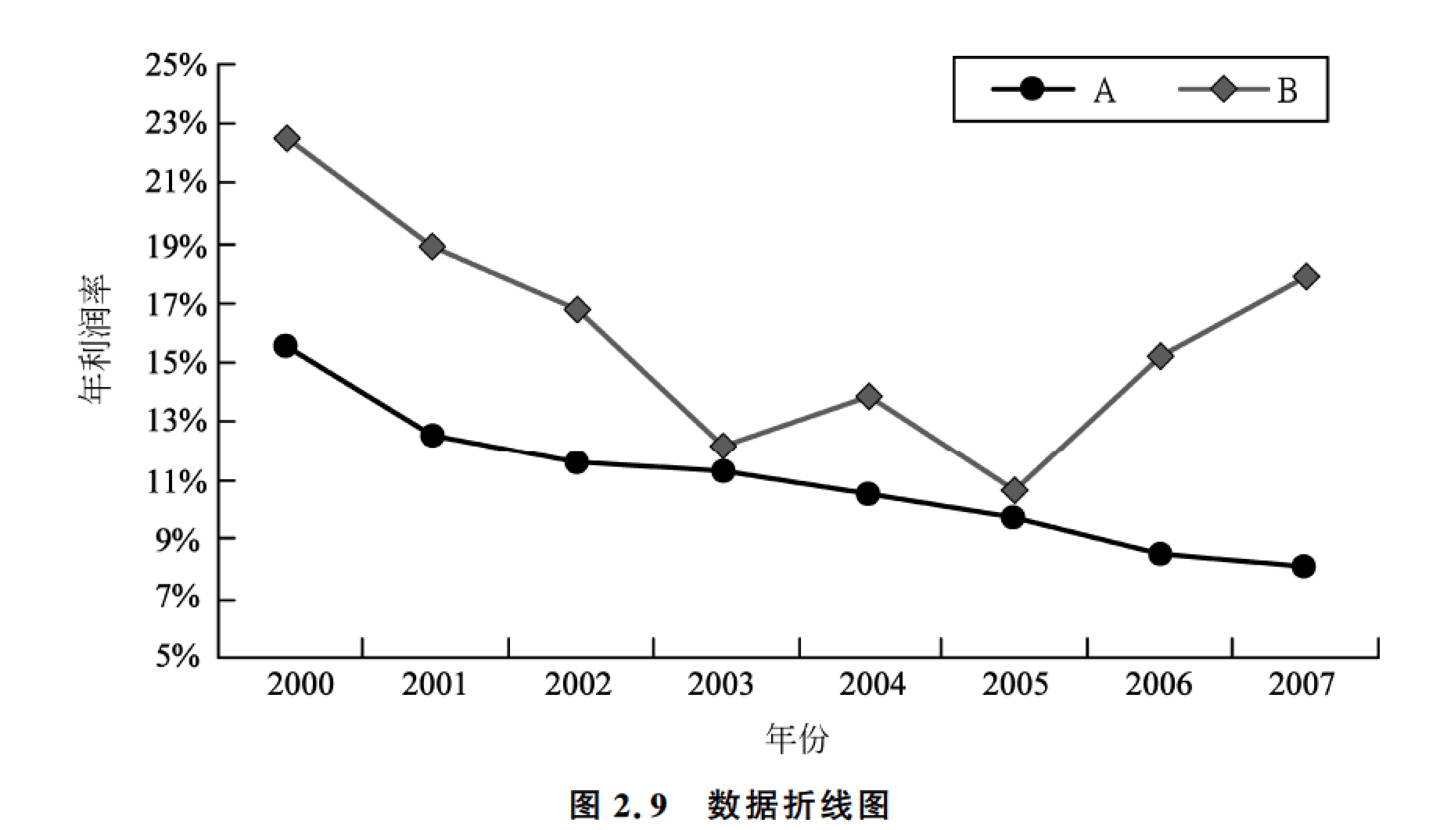


图 2.9 数据折线图

3. 散布图

散布图(scatter plot)是在 p 维空间中给出 p 个变量关系的点，借由点的疏密程度和延展方向等分布特征，通过图形观察数据，了解变量间的相互影响关系。由于维度过高的数据不易比较，因此散布图的维度通常等于或低于三维空间。回归分析中，常以散布图作为筛选独立变量 x 的基本检验步骤，通过绘制独立变量 x 与相依变量 y 的二维散布图，能初步得知 x 与 y 的相关性。图 2.10 给出了不同线性相关程度下的二维散布图。

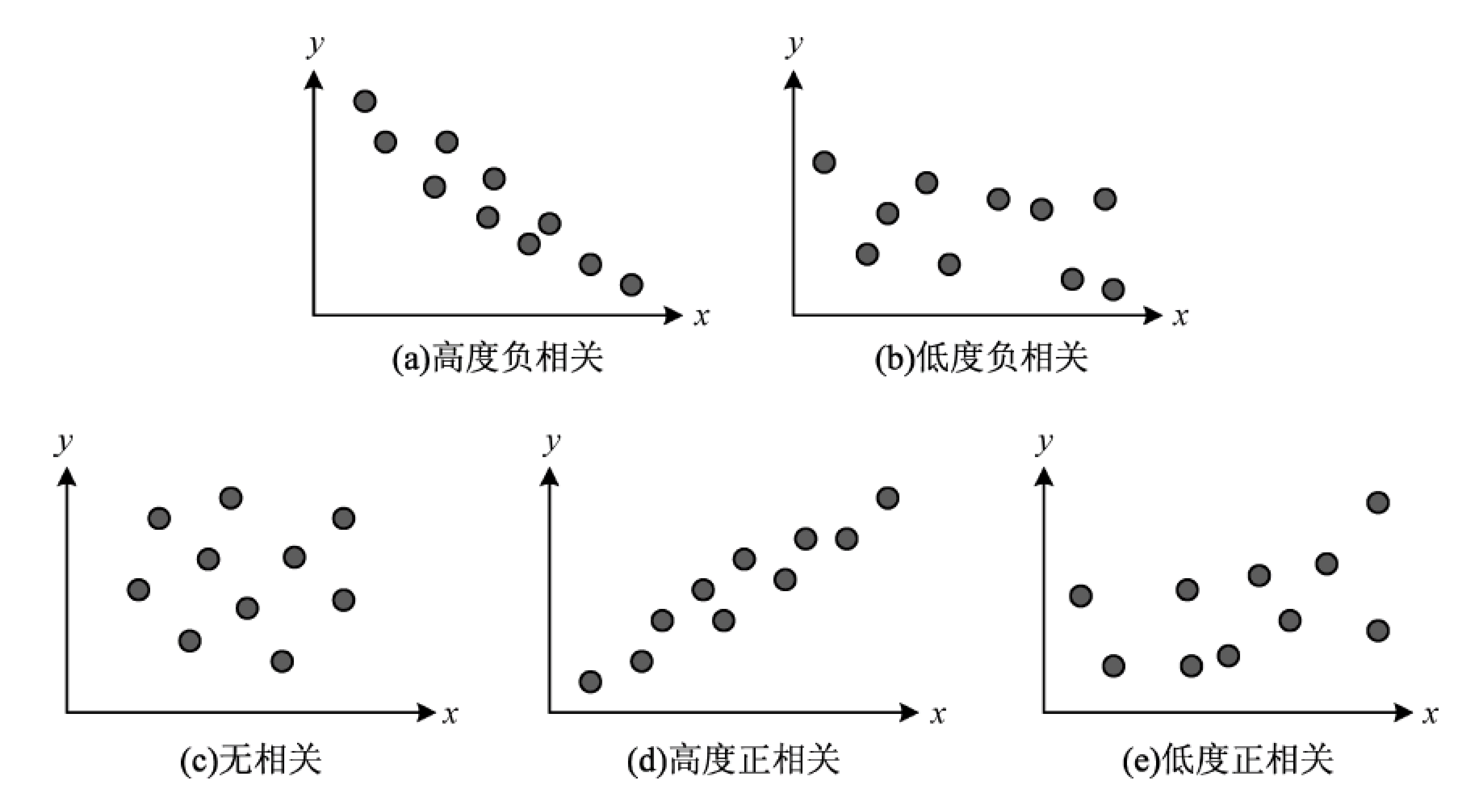


图 2.10　各种 x、y 相关程度所对应的散布图

4. 平行坐标图

平行坐标图(parallel coordinate plot)是一种用来检查高维度数据概况的图形呈现方法。打破传统散布图因受到坐标必须相互垂直的概念限制至多只能呈现三个维度的数据,改以平行坐标来呈现数据,使数据呈现不再受限于三个维度以内。平行坐标图是指在一份数据中,以 p 条垂直以及相互平行的坐标轴(坐标轴之间通常等距)来表示彼此之间不同的维度,每一笔数据以一条折线来呈现,折线与平行轴的相交位置为该数据于该维度变量所对应的数值。

在垂直坐标的显示中,各变量所对应的坐标轴均相互垂直,因此变量的顺序并不会对图形的呈现造成影响;但在平行坐标系统中,各坐标轴皆相互平行,轴与轴之间存在绝对的顺序关系。在实际应用中,分析者可概略检查坐标轴相邻的变量的相关性。以图 2.11 为例,该数据集为 50 个年龄在 20～33 岁的受访者数据,所搜集的变量为年龄、体重、身高与 BMI(body mass index)数值。图 2.11(a)以体重、年龄、身高、BMI 为变量顺序进行作图,以检查 50 位受访者各项特征的分布状态。由图可知男性受访者的平均体重与身高皆高于女性受访者。此外,也可观察到有一名男性受访者的身高体重皆异(高)于常人,另外也有一名男性受访者的 BMI 指数远高于其他受访者。

使用同样的数据,图 2.11(b)以年龄、体重、身高、BMI 为顺序进行作图。除了图(a)所能观察到的现象以外,在图(b)还可观察出身高与体重之间存在正相关;平均而言,身高越高体重也越重。相较于图(a),图(b)能够额外提供身高与体重间具相关性的信息,主要的关键在于图(b)的身高与体重为两相邻的坐标轴,因此能够呈现其相互关系;在图(a)中,因为身高轴与体重轴之间多了一个跟此二变量皆不相关的年龄轴,而无法观察出这些变量之间的相关性。因此,利用平行坐标图检查数据时,应尽可能将具相关性的变量摆放在相邻的坐标轴上,以加强图标所能提供的信息,必要时亦可尝试使用各种变量顺序来作图。图 2.11(c)与图 2.11(d)为平行坐标图的相关进阶应用,其分别撷取体重落于[50,54]与[68,72]的受访者数据来作图。从中可观察到若将体重限制于某段小范围时,身高与 BMI 之间的线条呈现类似对偶的性质,也就是说,当体重固定时,身高与 BMI 之间为负相关。

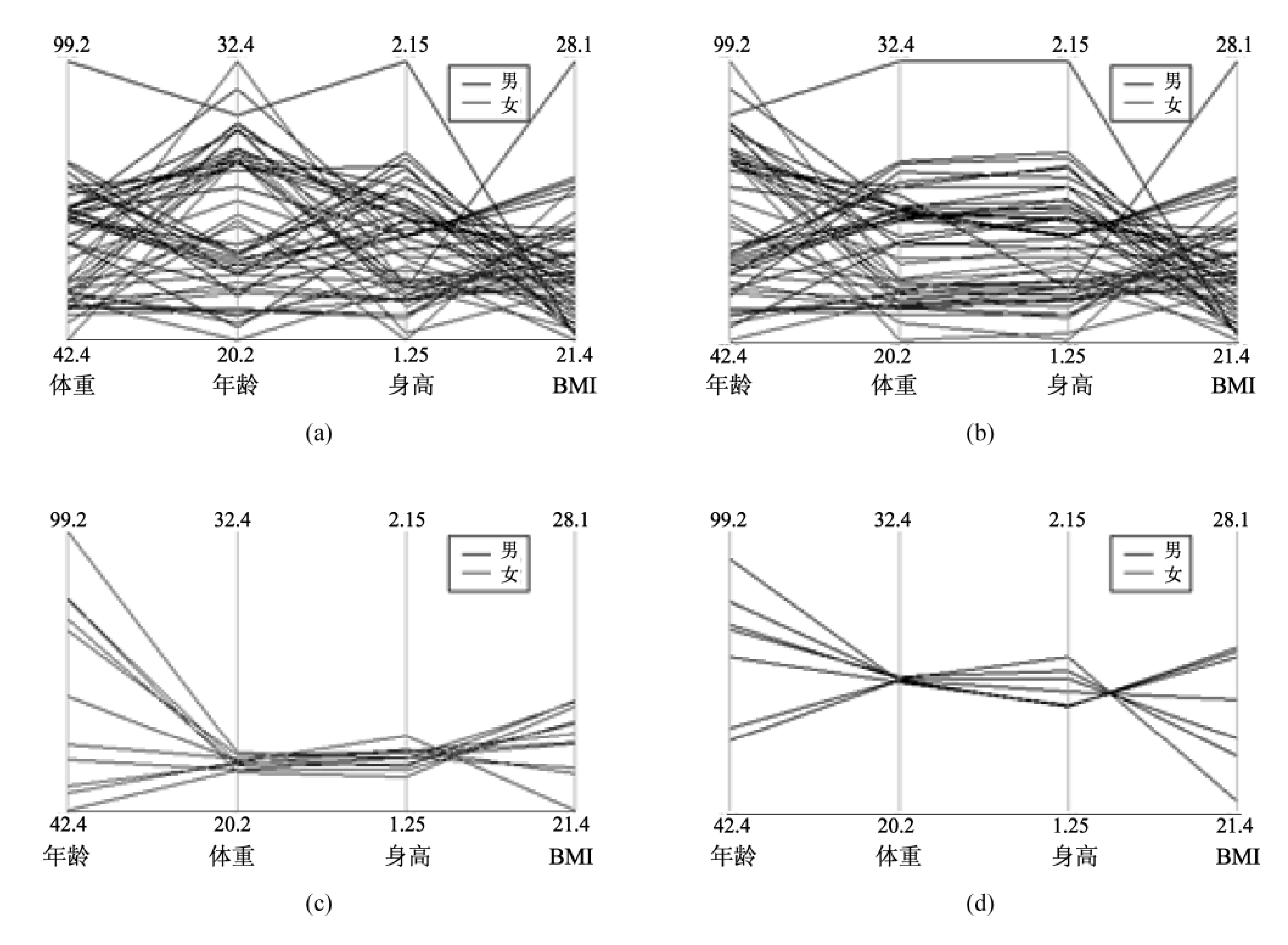


图 2.11 平行坐标图

2.7 数据整合与清理

由于人为疏忽、设备异常或抽样方法等因素，往往会发生数据误植、数据遗失或数据不一致、重复、矛盾等不同类型的数据问题，如表 2.2。若直接分析这些有问题的数据将会产生错误或无意义的结果，因此，必须借由数据整合与清理的过程，在建立数据挖掘模式前予以修正。以下将说明如何针对不同的数据问题进行数据预处理。

表 2.2 数据整合与清理的问题

问　　题	原　　因	数据准备步骤
不正确的数据	数据的值超出合理范围	数据整合
不一致的数据	不同源数据整合后所出现的分歧	
重复的数据	重复记录的字段或数值	
冗余的数据	出现相同意义的数据或字段	
遗漏值	测量设备或人为因素所造成的数据遗漏	数据清理
噪声	数据本身的误差或数据输入的偏误	
离群值	数据本身的特性、不当测量或数据输入错误	
数据尺度不合适	数据格式不符合挖掘工具的假设	数据转换
数据太多	数据或维度过高	数据归约

建立数据挖掘模式前,必须先将不同来源的数据汇整与分析成数据集,其来源可能是文件,如电子表格文件、文本文件,或在线数据库中的某一段数据表格,也可能是数据仓储中的数据方块。**数据整合(data integration)**的主要目的就是在解决多重数据储存(data store)或合并时所产生的数据不一致、数据重复或数据冗余的问题,以提高后续数据挖掘的精确度和速度。**数据清理(data cleaning)**的主要目的为填充或删除遗漏值、降低噪声与处理离群值数据。不同数据问题的处理方式,说明如下。

1. 不正确的数据

数据整合必须先确认数据的正确性与完整性,避免数据缺失造成结果的偏差。首先要确认数据的有效范围,例如一批晶圆的数量若不超过25片,机台的压力值不会产生负值。其次也要验证数据的合理性,例如某位学生的身高达1050cm,即可归类为不正确的数据。

2. 不一致的数据

数据不一致的处理是先修正不一致的记录,避免整合后的数据错误造成后续分析结果的误差。数据不一致的问题主要是由于整合数据后,不同来源的数据的属性可能不同,在数据表达、比例定义或编码上也会有所不同,因此产生数值或字段不一致的状况。针对数值的不一致,例如,重量属性在一个系统中可能是以公制的单位存放,而在另一个系统中则以英制的单位存放,此种单位差异可经由换算将其统一;若是数据内容本身的不同,例如,同一片晶圆在系统A记录的不良晶粒个数是10,在系统B记录的不良个数是5,则需进一步判定与检查以修正其中一笔数据。而字段的不一致,多半是属性命名不一致所造成。例如顾客姓名与会员姓名的域名虽然不同,但实际上填入的数据却是相同的,可将其中一个字段修正统一。

3. 重复的数据

数据重复的处理主要是针对重复出现的数值或字段。整合过后的数据常常会发生数据重复的问题,例如整合过后的数据表A中有机台的过站时间,在数据表B中也记录了机台的过站时间,若两项重复数据完全相同,则可选择删除其中一组记录,否则应注意哪一项记录为最新数据。

4. 冗余的数据

数据冗余的处理主要是针对具有相同意义或彼此间存有已知数学关系的字段,也就是此变量的属性或代表的意义可由另一变量推导而得。举例来说,若"年薪"可由"月薪"加"奖金"导出,则年薪就是多余的数据,可将年薪字段剔除。另外,属性命名的不一致,也有可能导致数据集中的冗余数据。

5. 遗漏值

遗漏值(missing value)为遗漏或错误的数据,可删除该笔数据或以特殊的方式补值。相较之下,空白值(empty value)为无法或不需填入的数据,例如,问卷调查允许某些人无须填入数据或跳题作答。数据遗漏可能包括人为或计算机数据输入的错误,输入时理解错误或认为不重要而没有输入,也有可能是搜集数据的设备出了问题,转换文件时出了问题,造成数据遗失。例如,测量机台故障,无法实时记录晶圆的过站时间。

有时遗漏值出现的样型本身就有意义,特别是问卷数据可能会反映"难言之隐",例如,

问卷调查时，当应答者不愿意回答年龄、年收入等问题时，即造成遗漏值的产生。

在数据搜集时，测量设备故障或人为因素造成的数据遗漏难以避免，所以必须在事后进行数据清理，降低数据遗漏对后续数据分析结果的影响。以下为几种处理遗漏数据的方法。

(1) **直接删去该变量值**：此为最直接简单的处理方法。然而，除非变量的属性有多个遗漏值，否则此方法并不奏效。但当数据遗漏比例很大时，此方法将造成大量数据流失。

(2) **人工填写遗漏值**：此方法费时且需额外增加人力成本，当数据集很大、遗漏值很多时，并不适当。

(3) **使用一个全局常数填充遗漏值**：将遗失的属性值用同一常数替换，如用无穷大符号"∞"替换"Unknown"，以符合后续分析的输入条件。此方法的缺点是仍无法解读遗失属性所隐含的信息。

(4) **使用属性平均值**：用该字段所有数据的平均值取代遗漏值。如用小学全校身高平均值替换身高属性中的遗漏值。缺点是不具客观性，当数据本身具有类别或等级之分时，容易高估或低估数据。

(5) **给定属于同一类别的所有样本的平均值**：利用具有相同等级或类别的数据平均值取代遗漏值。如利用全校六年级学生的平均身高来取代六年级学生遗漏的身高数据。

(6) **利用数据挖掘模式来填充遗漏值**：可用回归分析、决策树、人工神经网络等数据挖掘推导工具，详细方法将于后续各章陆续介绍。

不论用哪种模式来估计并补值，其目的都在于找到合理的替代值。在处理或取代遗漏值时可能会产生失真或误差的情况，例如，某些数据挖掘的方法可能无法处理遗漏值，因此在分析过程中必须删除整笔数据。或者，有些数据挖掘工具会用默认值取代遗漏值，导致失真的风险。此外，不同的填补方法对于挖掘结果的解释会有不同的影响，数据挖掘者必须清楚地了解每种取代方法的特性，才不会忽略原本应有的信息。

表 2.3(a)假设有一笔数据，调查 A～F 六位顾客的购买反应，问题包含了性别、年龄、薪水等信息，其中 F 顾客的购买反应为遗漏值，需进行补值。首先，将顾客 F 的数值以 0 表示，然后依照原始数据进行距离大小的比较，从中可以发现 F 顾客与 A～E 五位顾客的年龄、性别、薪水等距离如表 2.3(b)。接着再将 A～E 五位顾客的分数加总并且排序，得到 D 顾客为首要排序，接着对照 D 顾客的购买反应发现是 Yes，所以得出顾客 F 的购买反应为 Yes。

表 2.3 顾客基本数据

顾客基本数据(a)

顾客	性别	年龄	薪 水	购买反应
A	女	27	$ 19 000	No
B	男	51	$ 64 000	Yes
C	男	52	$ 105 000	Yes
D	女	33	$ 55 000	Yes
E	男	45	$ 45 000	No
F	女	45	$ 100 000	?

顾客基本数据(b)　　　续表

顾客	$d_{年龄_norm}$	$d_{性别_norm}$	$d_{薪水_norm}$	加总	由小到大排序	购买反应
A	1	0	1	2	5	No
B	0.33	1	0.44	1.77	4	Yes
C	0.38	1	0.06	1.44	2	Yes
D	0.66	0	0.55	1.21	1	Yes
E	0	1	0.67	1.67	3	No
F	0	0	0	0	—	**Yes**

6. 噪声

噪声(noise)表示一个数据中的随机误差或干扰。在数据输入时可能因人为因素或机器设备产生误差,而数据本身也可能存在随机误差,例如机台传感器故障,或是错误的数据传输以致搜集到不当的数据等。噪声的存在会造成有偏误的数据挖掘结果,导致结果的误判。针对噪声数据,若非数据本身存在的误差,经由噪声辨识后即可去除,若是数据本身既有的随机误差,可利用以下几种数据平滑(smooth)技术降低其对结果的影响。

(1) **分箱法**

分箱法(binning)的概念是利用“相邻”值来局部平滑储存在同一箱子的数据值。将数据排序后,依序排入预定的箱子中,排入方式可采用等宽(equal-width)或等深(频)(equal-frequency)方法,接着利用各箱子的平均值、中位数、边界值等三种数值进行数据平滑。

等宽分箱法是依照数据的数值范围来切割数据箱的间距,每一个分割的区间间隔相同,假设 X 和 Y 分别为该属性数据的最大和最小值,若将数据划分为 M 个区间,则可定义区间宽度为 $W=(X-Y)/M$。**等深分箱法**利用数据个数划分数据箱的区间,而每一个区间内的数据数相同,和等宽分箱法不同的是,其是将数据等分为数个数据箱,并经排序后,直接将数据装入所欲划分的 M 个区间。举例说明,假设欲分析 15 件商品的库存量,其数值依序分别是 5、10、12、12、24、32、43、55、60、65、72、77、81、90、120。为降低数据噪声,可将数据分为 5 个箱子,首先采用等宽分箱法,最大值和最小值分别为 120 和 5,因此间距为(120－5)/5＝23,所以各箱子之间的宽度为 23。因此第一个箱子内的库存数据为 5、10、12、12、24,第二个箱子为 32、43,第三个箱子为 55、60、65、72,第四个箱子为 77、81、90,而第五个箱子则只装一个数值 120,如图 2.12。



图 2.12　等宽分箱法

若采用等深分箱法,同样分割为 5 个箱子,每个箱子装 3 个数值,第一个箱子是 5、10、12,第二个箱子是 12、24、32,第三个箱子是 43、55、60,第四个箱子是 65、72、77,第五个箱子

则是 81、90、120，如图 2.13。



图 2.13　等深分箱法

（2）**数据配适**

利用数据配适为新的函数来平滑数据，例如采用简单线性回归以一个解释变量估计目标变量，详细回归方法第 9 章会进一步介绍。

7. 离群值

在搜集的数据中，若某一些数据的表现明显与其他数据不一样时，这些数据称为离群值，例如，某班同学的身高大都集中在 150～160cm，但有某几位同学身高超过 200cm，则称这些同学的身高是离群值。离群值会影响挖掘模式的效果，特别是预测模式，因此，在建立挖掘模式前必须先行处理离群值，主要有以下三种处理方法。

（1）**直接删除**

当发现数据是出自于仪器或工具造成的判断错误，或者是数据完全不合理的时候，即可考虑直接删除该笔数据。

（2）**用其他数值替换，将数据范围归一化**

当数值变量为空白值或是非数值数据，且数据具有一定的代表性时，则可以其他数值来做更替，将数据的范围归一化，例如以 0 与 1 来表示，归一化方法参考 2.8.1 节。

（3）**聚类分析**

离群值可利用聚类分析检测而得，借由将类似的点结合为一个群组或族群，落在聚类集合之外的值即视为离群值，关于聚类分析详细内容可参考第 6 章。

若与领域专家进行讨论后，该离群值存有特殊意义或为分析的主要目的，则予以保留。例如，对信用卡从业者而言，每月使用且刷卡额达数百万金额的顾客虽为少数，却是重要的黄金客户，此笔具有特殊意义的数据即可保留；反之，若无特殊意义，则可直接删除。

2.8 数据转换

数据转换（data transformation）为将数据转换成适合数据挖掘模式可处理的数据格式或为丰富化数据的内容，以转换原始数据或重新编码以提升数据价值，其中可能涉及数据数值与数据类别的转换。例如，将数值型数据转换为离散型的类别数据，根据领域知识将旧有变量合并成新的变量，亦或将数据归一化以避免尺度的差异，常见如人工神经网络对输入数据的归一化。

2.8.1 数据数值转换

1. 归一化

归一化（normalization）是将属性数据按比例缩放到一个特定的区间，如[－1，1]或[0，1]。例如人工神经网络中的反向传播（back propagation）算法需要对于训练样本输入值

范围转换至[0,1]。归一化可防止较大初始值域与较小初始值域属性间互相比较的情况，以及权重过大的问题。

极小值—极大值归一化(min-max normalization)是常用的归一化方法，主要是对原始数据进行线性转换，假设 X_A^{min} 和 X_A^{max} 分别为属性 A 的最小值和最大值。其计算如式(2.1)所示：

$$X' = \frac{X - X_A^{min}}{X_A^{max} - X_A^{min}}(X_{A,new}^{max} - X_{A,new}^{min}) + X_{A,new}^{min} \tag{2.1}$$

将 A 的值输入到区间 $X_{A,new}^{max} - X_{A,new}^{min}$ 中得到 X'。极小值—极大值归一化应保持原始数据值之间的关系。如果输入的值落在 A 的原始数据区之外，将产生超出范围的错误。

例如，假设属性收入的最小与最大值分别为＄15 000 和＄95 000，若想要将收入转换到区间[0,1]。根据极小值—极大值归一化的方法，收入值＄73 500 将转换为

$$X' = \frac{73\,500 - 15\,000}{95\,000 - 15\,000}(1 - 0) + 0 = 0.731\,25$$

2. 标准化

数据标准化(standardization)是基于属性 A 的平均值和属性 A 的标准差将数据标准化。A 的值 X 标准化后为 Z，可经由式(2.2)计算而得

$$Z = \frac{X - \overline{X}_A}{S_A} \tag{2.2}$$

其中，$\overline{X}_A$ 与 S_A 分别为属性 A 的平均值和标准差，当属性 A 的最大值和最小值未知，或孤立点左右极小值—极大值归一化时，可改用标准化方法。

例如，假设属性收入的平均值与标准差分别为＄55 000 和＄15 000。以式(2.2)进行标准化后，收入值＄73 500 将转换为

$$\frac{73\,500 - 55\,000}{15\,000} = 1.233$$

2.8.2　数据属性转换

1. 离散型数据转成连续型数据

离散型数据转换成连续型数据必须加入领域知识来定义离散值的距离或相似程度。此过程通常需要结合专家意见，然后以类似的矩阵定义出数值与数值之间的距离或相似程度，再利用此距离或是相似程度把离散的数据转换为连续型的数据形态。例如，学生成绩的等级为 A 应该对应至 85 分，若成绩为 B+，则应该对应至 78 分。

2. 连续型数据转成离散型数据

离散化(discretization)是将连续数据分布到数个小区间，以类别尺度取代原有连续数据的尺度。经由离散化后的数据在叙述上较为简单，可使通过数据挖掘或机器学习方法所得到的结果更容易被了解与解释(Liu *et al*.,2002)。离散化的区间切割不足会造成准确度降低或解释能力下降，而区间切割太多则会失去离散化的意义。数据在离散化后，原有的信息多少会有所遗失，但不当的离散化方法可能造成信息的大量遗失或提供不正确的信息。

典型离散化的过程包含四个步骤：①将欲转换的连续数值排序；②选择分割或合并的准则；③分割或合并数值；④是否符合停止条件。

数据离散化可同时进行特征的选择与数据维度化约。有些方法需要类别信息，有些则不用，分箱法为简单常用的离散化方法，除此之外，还有利用熵(entropy)尺度进行二维分支的 ID3(Quinlan，1986)与 C4.5(Quinlan，1993)等决策树方法，详细内容可参考第 4 章，或利用聚类分析将数据分成几个群组，每一群组即可代表一个区间，并将数据归属于对应的区间以进行离散化，聚类分析详细内容可见本书第 6 章。其他具代表性的离散化方法还有使用二位递归分支算法的 D2(Catlett，1991)、使用最小叙述长度准则法(minimum description length principle，MDLP)来改善 D2 无限递归分支的缺点(Fayyad & Irani，1993)、使用 Mantaras 距离进行离散化(Cerquides & de Mantaras，1997)、一层离散分支的 1R 分类算法(Holte，1993)以及关联性作为衡量两连续变量相依程度的 Zeta 离散法(Ho & Scott，1997)。

2.9 数据归约

数据本身的价值因数据分辨率(resolution)的不同而有所差别，例如年、季、月、星期等对数据代表的意义与信息亦不尽相同。可经由数据汇总(aggregation)以提升数据代表的意义，例如，计算销售数据时，可先集中计算日销售数据，再计算月和年的销售额。在分析过程中，数据集的大小与数据的分布差异皆会影响挖掘效果，例如某一类型的数据特别稀少，容易造成分类模型忽略该类型数据，造成挖掘的结果偏离所关心的目标。

在数据搜集阶段，应尽可能地搜集所有可记录的变量或数据，以免遗漏对目标变量具有潜在影响的变量或数据。搜集而来的原始数据必须再经由数据归约，删除或过滤数据集合中不具代表性或无用的数据，以减少数据挖掘的时间与成本，获得更具利用价值的数据。亦即数据归约的主要目的是得到与原始数据具有相同信息但却较精简的数据集，并具有以下效益：

(1) 提升数据质量：精简后的数据与原始数据虽有差异，但对欲提取的信息准确性与代表性并不一定较差，反而有助于提高知识的应用性以及准确性，并且降低无用以及错误数据的影响，提升数据质量。

(2) 缩短数据挖掘时间：数据挖掘的数据量越多，所需的处理时间也越长。因此，可选择少量具代表性的数据以加快数据处理速度。

(3) 简单的规则，有助于数据价值的提升、知识价值的取得与增加可读性。

(4) 降低数据储存成本：使后续的数据搜集仅需搜集缩减后的数据集合。

数据集合是指数据集或数据库中的数据表。数据表中描述数据集合所用的特征或属性称为数据维度(dimension)，根据数据维度所描述的数据集合称为数据记录，记录数据集合于某一维度下的数值称为数据数值(value)，在某一维度下所有可能出现的数值称为值域(domain)。数据维度归约可以减少数据记录的长度，数据记录归约能够减少数据记录的笔数，而数据数值归约则能缩小可能的值域。以下则分别针对数据维度归约、数据数值归约进行说明。

2.9.1 数据维度归约

数据维度归约常用在分类或预测的问题。最直接的方式是以目标变量作为比较基准，

利用特征选取法将变量维度与目标变量不相关的属性删除。另一个方法是利用主成分分析法将变量作线性转换,只留下提供较多信息的几个主成分,借以缩小变数维度。此法不需要目标变量作为比较基准,目的在于找出最能解释数据变异的线性组合。

1. 特征选取法

所谓特征选取(feature selection)是依据所规定的特征衡量条件,删除不相关的特征或属性,以选取用于分析数据的最佳特征的过程(Liu & Motoda,1998)。其操作步骤依序为:决定特征衡量准则、选取特征产生计划、选定搜索策略、设定停止条件。

以下以制程加工时间数据表为例说明特征选取法的应用。首先,假设制程"加工时间≤30"者标识为类别 1,"30<加工时间≤40"者标识为类别 2,"加工时间>40"者标识为类别 3,则表 2.4 可转换为表 2.5。

表 2.4 制造数据表

产品编号	制程 A		制程 B		制程良率
	加工时间/min	机台类型	加工时间/min	机台类型	
01	28	A01	48	B03	0.53
02	27	A01	42	B03	0.62
03	31	A03	43	B01	0.84
04	42	A02	33	B02	0.91
05	46	A02	28	B03	0.85
06	50	A01	27	B03	0.68
07	35	A02	24	B01	0.83
08	24	A03	36	B02	0.69
09	28	A02	25	B01	0.88
10	44	A03	37	B03	0.92

表 2.5 离散化后的制造数据表

产品编号	制程 A		制程 B		产品制程良率
	加工时间/min	机台类型	加工时间/min	机台类型	
01	1	A01	3	B03	低
02	1	A01	3	B03	低
03	2	A03	3	B01	高
04	3	A02	2	B02	高
05	3	A02	1	B03	高
06	3	A01	1	B03	低
07	2	A02	1	B01	高
08	1	A03	2	B02	低
09	1	A02	1	B01	高
10	3	A03	2	B03	高

步骤一：决定特征衡量准则。在此先介绍四种常见衡量数据维度的方法及其应用：

(1) **一致性测量法(consistency measurement)**

假设 C(制程 A 加工时间,制程良率)表示制程为 A 加工时间对制程良率具有不一致数据数值的笔数,而 C(制程 A 加工时间$_{(i)}$,制程良率)表示制程 A 加工时间为第 i 类时,会造成制程良率不一致的笔数。则可计算制程 A 加工时间、制程 A 机台类型、制程 B 加工时间、制程 B 机台类型四个特征所产生不一致的数据笔数如式(2.3)：

$$C(X,Y)=\sum_{i=1}^{n}C(X_{(i)},Y) \tag{2.3}$$

C(制程 A 加工时间,制程良率)

$=C$(制程 A 加工时间$_{(1)}$,制程良率)$+C$(制程 A 加工时间$_{(2)}$,制程良率)

$+C$(制程 A 加工时间$_{(3)}$,制程良率)

$=1+0+1=2$

C(制程 A 机台类型,制程良率)

$=C$(制程 A 机台类型$_{(A01)}$,制程良率)$+C$(制程 A 机台类型$_{(A02)}$,制程良率)

$+C$(制程 A 机台类型$_{(A03)}$,制程良率)

$=0+0+1=1$

C(制程 B 加工时间,制程良率)

$=C$(制程 B 加工时间$_{(1)}$,制程良率)$+C$(制程 B 加工时间$_{(2)}$,制程良率)

$+C$(制程 B 加工时间$_{(3)}$,制程良率)

$=1+1+1=3$

C(制程 B 机台类型,制程良率)

$=C$(制程 B 机台类型$_{(B01)}$,制程良率)$+C$(制程 B 机台类型$_{(B02)}$,制程良率)

$+C$(制程 B 机台类型$_{(B03)}$,制程良率)

$=0+1+2=3$

由以上计算可得知制程 A 机台类型对制程良率所产生数据维度不一致的笔数最低,故与其他变量比较,制程 A 机台类型对制程良率有明显的区分。

(2) **关联性测量法(association measurement)**

假设 R(制程 A 加工时间,制程良率)表示制程 A 加工时间对制程良率相关联的程度,而 R(制程 A 加工时间$_{(i)}$,制程良率)表示制程 A 加工时间为分类 i 与制程良率的关联程度。则制程 A 加工时间、制程 A 机台类型、制程 B 加工时间、制程 B 机台类型四个特征与制程良率的关联程度可计算如式(2.4)：

$$R(X,Y)=\prod_{i=1}^{n}R(X_{(i)},Y) \tag{2.4}$$

R(制程 A 加工时间,制程良率)

$=R$(制程 A 加工时间$_{(1)}$,制程良率)$\times R$(制程 A 加工时间$_{(2)}$,制程良率)

$\times R$(制程 A 加工时间$_{(3)}$,制程良率)

$=\frac{3}{4}\times 1\times\frac{3}{4}=\frac{9}{16}$

R(制程 A 机台类型,制程良率)

$= R$(制程 A 机台类型$_{(A01)}$,制程良率)$\times R$(制程 A 机台类型$_{(A02)}$,制程良率)

$\times R$(制程 A 机台类型$_{(A03)}$,制程良率)

$$= 1 \times 1 \times \frac{2}{3} = \frac{2}{3}$$

R(制程 B 加工时间,制程良率)

$= R$(制程 B 加工时间$_{(1)}$,制程良率)$\times R$(制程 B 加工时间$_{(2)}$,制程良率)

$\times R$(制程 B 加工时间$_{(3)}$,制程良率)

$$= \frac{3}{4} \times \frac{2}{3} \times \frac{2}{3} = \frac{1}{3}$$

R(制程 B 机台类型,制程良率)

$= R$(制程 B 机台类型$_{(B01)}$,制程良率)$\times R$(制程 B 机台类型$_{(B02)}$,制程良率)

$\times R$(制程 B 机台类型$_{(B03)}$,制程良率)

$$= 1 \times \frac{1}{2} \times \frac{3}{5} = \frac{3}{10}$$

由以上计算可得知制程 A 机台类型与目标变量制程良率的关联程度最高,故与其他变量比较,制程 A 机台类型对制程良率有明显的区分。

(3) **判别测量(discriminant measurement)**

假设 D(制程良率,制程 A 加工时间)表示制程 A 加工时间对制程良率能被正确判别的比率,而 D(制程良率,制程 A 加工时间$_{(j)}$)表示制程 A 加工时间为分类 j 时,对制程良率的鉴别能力。因此,制程 A 加工时间、制程 A 机台类型、制程 B 加工时间、制程 B 机台类型四个特征对制程良率的鉴别能力可计算如式(2.5):

$$D(Y,X) = \min\{D(Y_{(j)},X)\}, \quad j = 1,2,\cdots,m \tag{2.5}$$

D(制程良率,制程 A 加工时间)

$= \min\{D$(制程良率$_{(低)}$,制程 A 加工时间$),D$(制程良率$_{(高)}$,制程 A 加工时间$)\}$

$$= \min\left\{\frac{3}{4},\frac{1}{2}\right\} = \frac{1}{2}$$

D(制程良率,制程 A 机台类型)

$= \min\{D$(制程良率$_{(低)}$,制程 A 机台类型$),D$(制程良率$_{(高)}$,制程 A 机台类型$)\}$

$$= \min\left\{\frac{3}{4},\frac{1}{2}\right\} = \frac{1}{2}$$

D(制程良率,制程 B 加工时间)

$= \min\{D$(制程良率$_{(低)}$,制程 B 加工时间$),D$(制程良率$_{(高)}$,制程 B 加工时间$)\}$

$$= \min\left\{\frac{1}{2},\frac{1}{2}\right\} = \frac{1}{2}$$

D(制程良率,制程 B 机台类型)

$= \min\{D$(制程良率$_{(低)}$,制程 B 机台类型$),D$(制程良率$_{(高)}$,制程 B 机台类型$)\}$

$$= \min\left\{\frac{3}{4},\frac{1}{2}\right\} = \frac{1}{2}$$

由以上计算可得知制程 A 机台类型与制程良率的鉴别能力最高,故与其他变量比较,制程 A 机台类型对制程良率有明显的区分。

(4) 信息增益测量(information measurement)

又称决策树特征选取法,其目的是通过决策树的熵,衡量变量对目标变量的区分能力,去除较不相关或多余的变量,或是通过样本的选取技术删除数据库中重复以及错误的数据,详细内容可见第 4 章决策树分析。

步骤二:选取特征产生计划。在表 2.5 中,除了产品编号与产品制程良率外,还须考虑其余四个特征:制程 A 加工时间、制程 A 机台类型、制程 B 加工时间、制程 B 机台类型所有的特征晶格(lattice)组合,如图 2.14 所示。常见的特征产生计划方法有以下四种。

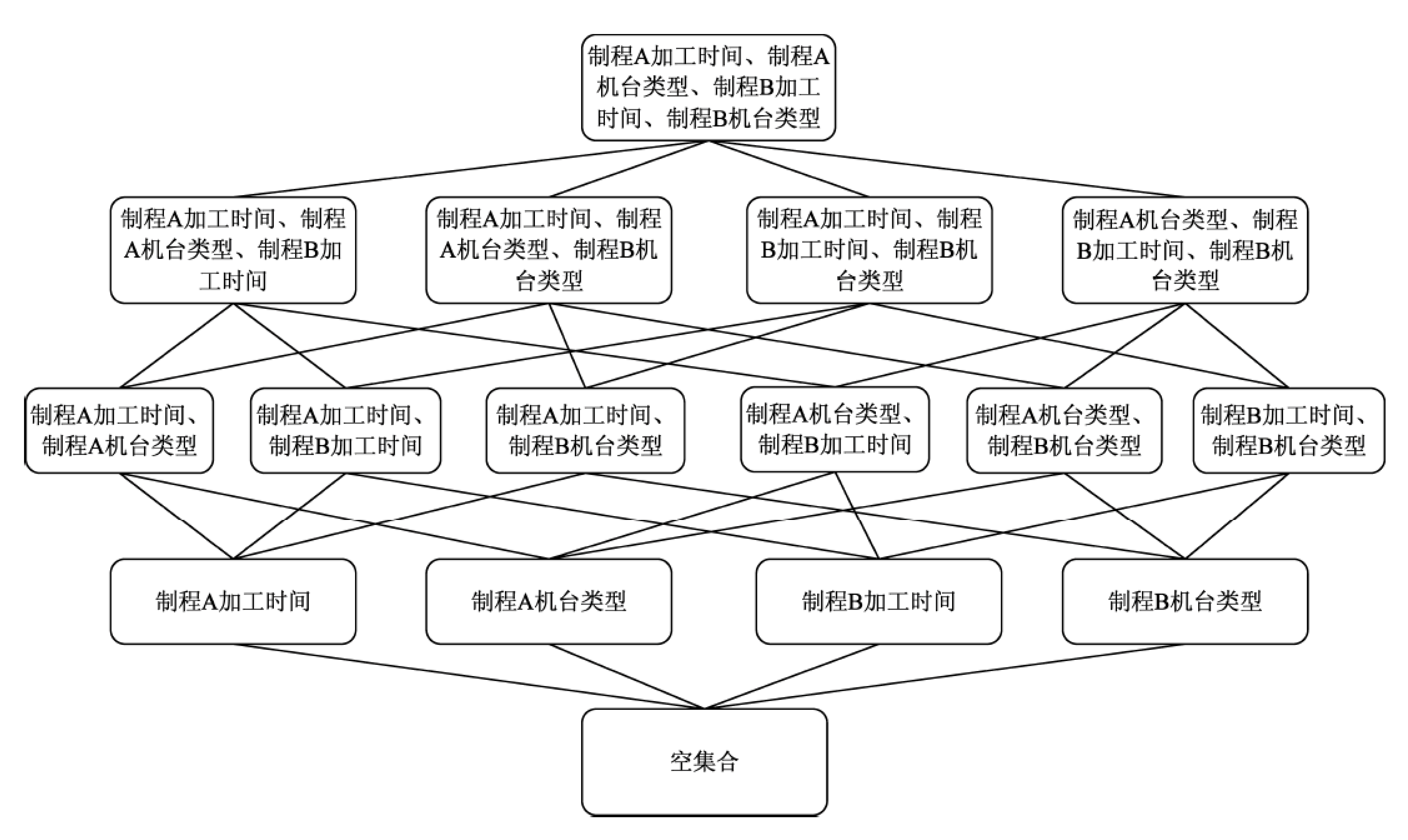


图 2.14 制程数据表的特征晶格组合

(1) 逐步向前挑选法(sequential forward generation)

逐步向前挑选法顺序的产生是由晶格下方到晶格上方,每次多考虑一个数据维度。首先依据所选择的测量法去计算第一层晶格的单一数据维度,并从中挑选出最好的数据维度,然后分别计算晶格第二层成对数据维度的测量值,最后选出最好的测量值以和之前最好的测量值比较,以此类推。

(2) 逐步向后删减法(sequential backward generation)

逐步向后删减法是由晶格上方往晶格下方中每次都少考虑一个数据维度。首先依据公式计算精简任一数据维度的可能组合,并且从中挑选最好的数据维度;接着针对晶格的单一数据维度分别计算测量值,最后再和之前的测量值相比较。

(3) 混合法(bidirectional generation)

混合法结合了逐步向前挑选法以及逐步向后删减法,同时从晶格下方的{ }(空集合)往晶格上方与晶格下方出发。

(4) 随机选取法(random generation)

随机选取法为配合随机列举策略衍生而来的方法。首先,以随机的方式决定由晶格上

方或晶格下方出发，配合随机列举策略去产生任何一种可能的数据维度组合，并进行审核。

步骤三：特征选取策略。特征选取策略取决于特征维度，假设数据中存有 N 个维度，所有可能的特征组合为 $2^N(2^N=C_0^N+C_1^N+C_2^N+\cdots+C_N^N)$，其中，2 的意思是选取或不选取这个特征。由此可知，特征选取策略的计算时间与空间取决于特征维度；当维度增加到数百甚至数千个时，数据维度归约所需的计算时间与成本将快速增长，使得此策略难以使用。因此，用户可考虑时间与成本自行规定停止条件，例如，不一致的数据笔数少于 3、信息增益大于 0.8、相关程度大于 95%、数据特征组合大于 5 等。以下将探讨经常采用的两种特征选取策略：穷举搜索策略与启发式搜索策略。

(1) **穷举搜索策略(exhaustive search strategy)**

穷举搜索策略是将所有可能的组合列出，比较不同特征维度，以找出最佳特征组合的策略，其采用先宽再深(breadth-first)的方式搜索每一层的组合，如图 2.15 所示。此方法虽然最简单，且能找出最佳的特征组合，但却非常耗时。然而，若选用单调的(monotonic)衡量基准，则可使用**完全搜索策略(complete search strategy)**，例如，分支界限法(branch and bound method)(Narendra & Fukunaga，1977)，不仅可减少搜索个数，还可保证能找到最佳特征组合。然而，在无法满足单调性的条件时，为了求得最佳组合，只能采用穷举搜索策略。

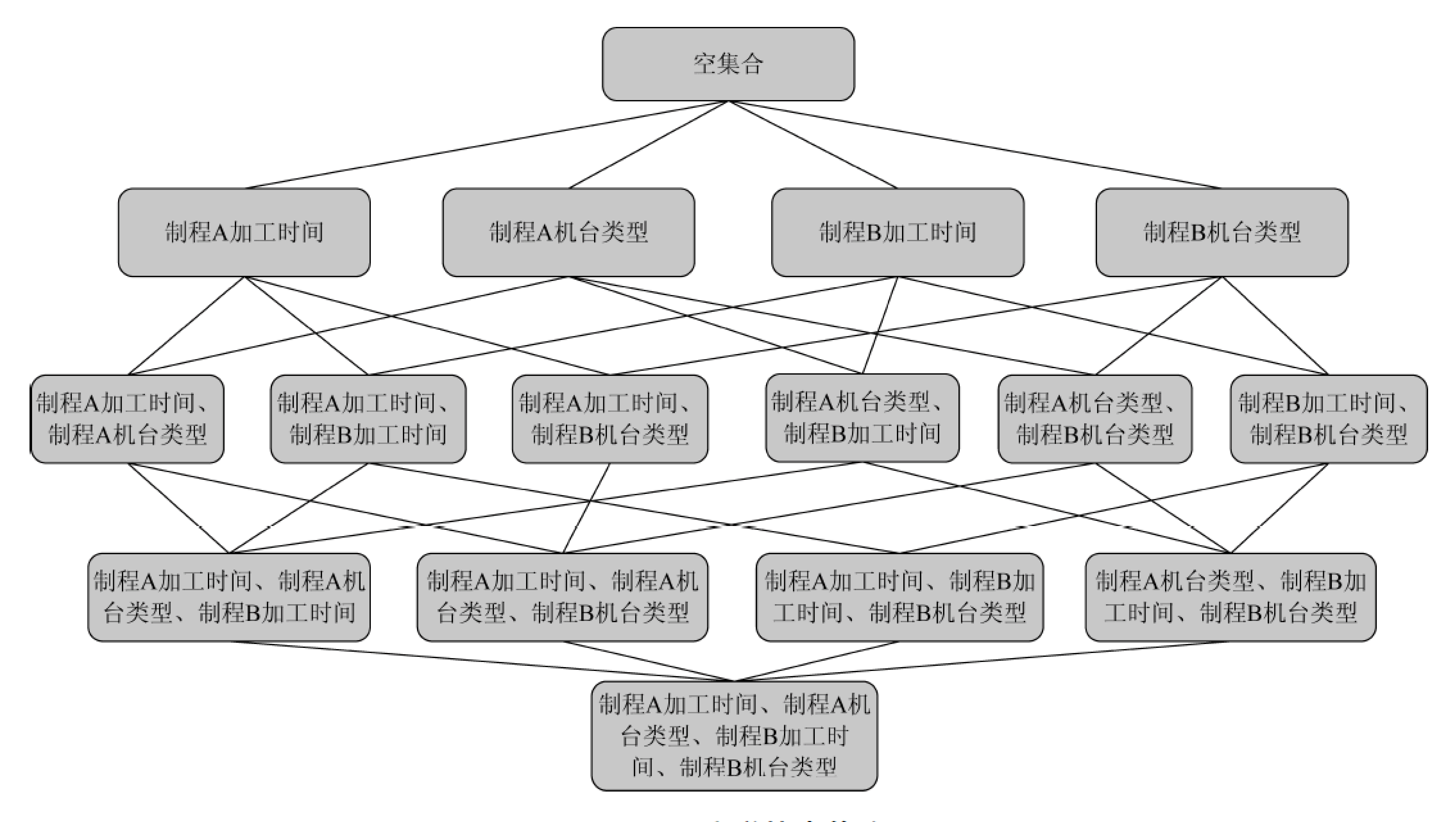


图 2.15　穷举搜索策略

(2) **启发式搜索策略(heuristic search strategy)**

启发式搜索策略可以利用贪婪的(greedy)方法，以所选的特征为基础，一步一步搜索。例如，深度优先搜索法(depth-first search)是先从各特征中选取 N 个最佳的特征，接着根据所选的特征产生 N 个维度的组合，并挑选最好的 N 个组合，以此类推。假设 $N=1$，若第一层所选取的特征是制程 A 加工时间，接着考虑包括制程 A 加工时间的组合，如图 2.16 所示。在搜索特征空间时，启发式搜索策略借由搜索局部最佳组合(灰底的部分)，达到与穷举搜索策略相去不远的特征组合。虽然不保证能得到最佳解，但有较高的执行效率。

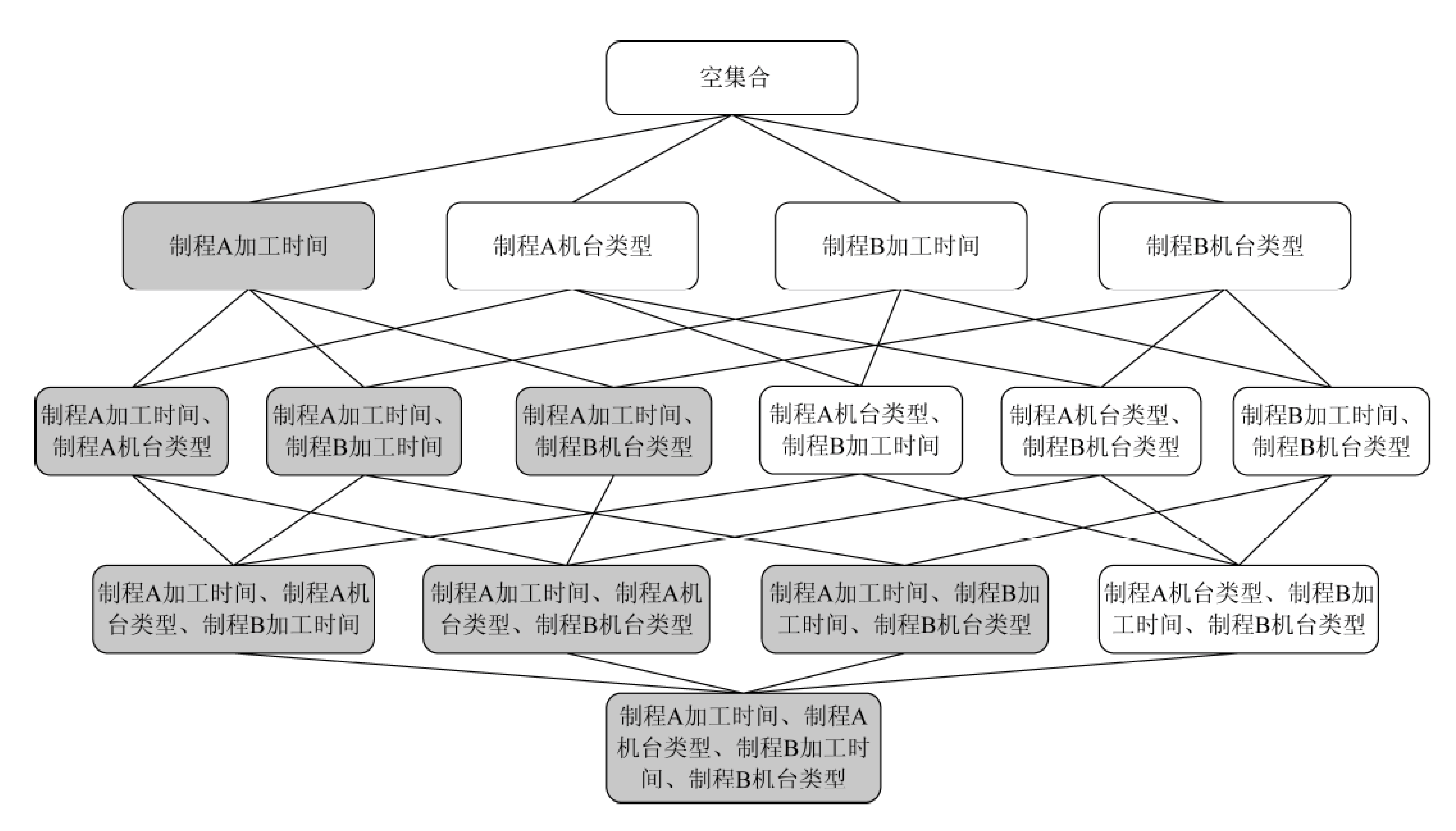


图 2.16 启发式搜索法所需计算的组合(以制程 A 加工时间为例)

(3) **随机搜索策略(random search strategy)**

随机搜索策略是以所选的特征为衡量基准,以随机增加或删除特征的方式,任意增删特征的维度,不断改进不同的特征组合以产生较佳的组合,直到符合所设定的停止条件。

步骤四:设定停止条件。当计算的选取属性子集合其衡量准则结果满足设定门槛,则停止,例如,一致性测量结果小于 2。由于此阶段的目的是进行数据归约,因此只要满足停止条件即可,不一定要找出最佳数据特征组合。

2. 主成分分析法

假设数据包括了 P 个属性的数值或是数据向量,主成分分析法(principal component analysis,PCA)是挑选最能表示数据变异的 k 个维度的正交向量 $k \leqslant P$,因而产生了维度的缩减。PCA 和直接剔除属性不同,其是将原始数据转换至另外几个主成分变量,亦即仍须输入其原始数据以产生新的主成分,因此仅是计算维度的减少,数据输入的维度则未改变。PCA 所产生的主成分分析可以当成是多元回归或是分群的输入。

2.9.2 数据数值归约

数据挖掘主要是找出较高层次的知识,如特殊的样型或趋势,以协助决策者制订方案,因此需将原始数据中太细或较低层次的数据离散化与广义化,使简化后的数据更有意义,且更容易解释,以利知识的取得与发掘,同时节省数据存放空间,增进挖掘效率。连续型数据可使用离散化方法,将属性值域分为若干区间,而离散型数据则可使用概念阶层。以下将分别对连续型数据与离散型数据的归约技术进行说明。

1. 离散化

有时离散型的数据比连续型的数据更容易解释。此时就必须将连续型数据离散化,以符合工具能处理的数据格式。在数值归约方面,通过将属性值域划分为区间范围,离散化技

术可以减少连续尺度值的数据个数(Han & Kamber,2011)。详细的离散化方式,参考2.8.2节数据属性转换的介绍。

2. 概念阶层

连续型数据数值具有大小顺序关系,通过离散化技术可将其划分为几个不同的区间。离散型数据数值因为本身往往仅具名目上的意义,并无法得知其数值是否相同或数值差异大小等,所以无法使用相同的方法达到数据数值归约的目的。而需使用概念阶层(concept hierarchy generation)将数据一般化(generalization),并用高阶层概念替换低阶层"原始"数据。例如:分类属性,如"街道",可以概化为较高层的概念,如"地区"或"城市";同样地,数值属性如"时间",可以映射到较高层的概念,如"天"、"周"、"月"、"季"和"年"。概念阶层的定义可由系统用户、领域专家等以人为方式主观规定,借由这些阶层的关系,将可有效厘清数据。

以表2.6来说,针对液晶面板尺寸的数据特征及产品所需尺寸大小,用户可将尺寸定义为大、中、小,其中手机、数码相机、掌上型电玩、电子字典所使用的是小尺寸面板;家电使用面板、车用液晶屏幕、笔记本电脑、工厂用设备操作屏幕所使用的是中尺寸面板;桌面计算机屏幕、数字电视、广告面板等则是使用大尺寸面板。由图2.17可知,从最高层的概念液晶面板,到最详细信息的手机、笔记本电脑及数字电视等原始概念,即是整个概念阶层的组合元素与架构,越上层的概念所包含的范围就越广,反之则越窄。

表2.6　面板商品

产品编号	液 晶 面 板	产品编号	液 晶 面 板
01	数字电视	07	广告面板
02	车用液晶屏幕	08	桌面计算机屏幕
03	掌上型电玩	09	工厂用设备操作屏幕
04	数码相机	10	笔记本电脑
05	家电使用面板	11	电子字典
06	手机		

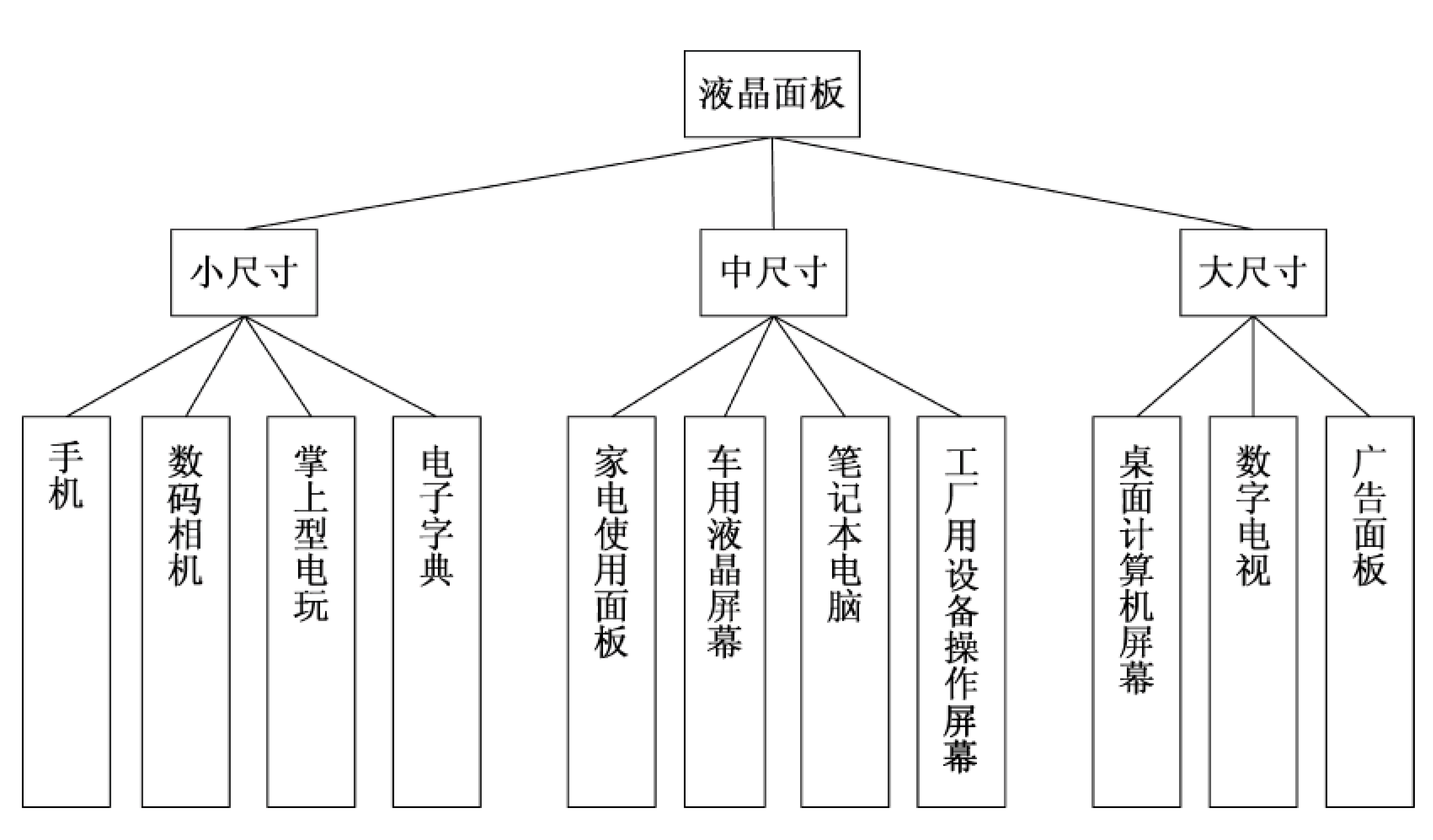


图2.17　TFT-LCD面板产品种类的概念阶层

2.10 数据分割

数据分割(data partition)是将数据分成训练数据组(training data)、测试数据组(testing data)、验证数据组(validation data),训练数据是用以建立模式,测试数据是用以评估训练数据所建立的模式是否过度复杂或其通用性,验证数据则是用以衡量模式的好坏,例如分类错误率(mis-classification rate)、均方误差(mean-squared error)。一个好的训练模式应该对于未知的数据仍保有很好的配适度,若当模式复杂度越来越高,而测试数据的误差却越来越大,表示该训练模型有过度配适(overfitting)的情形,如图 2.18。

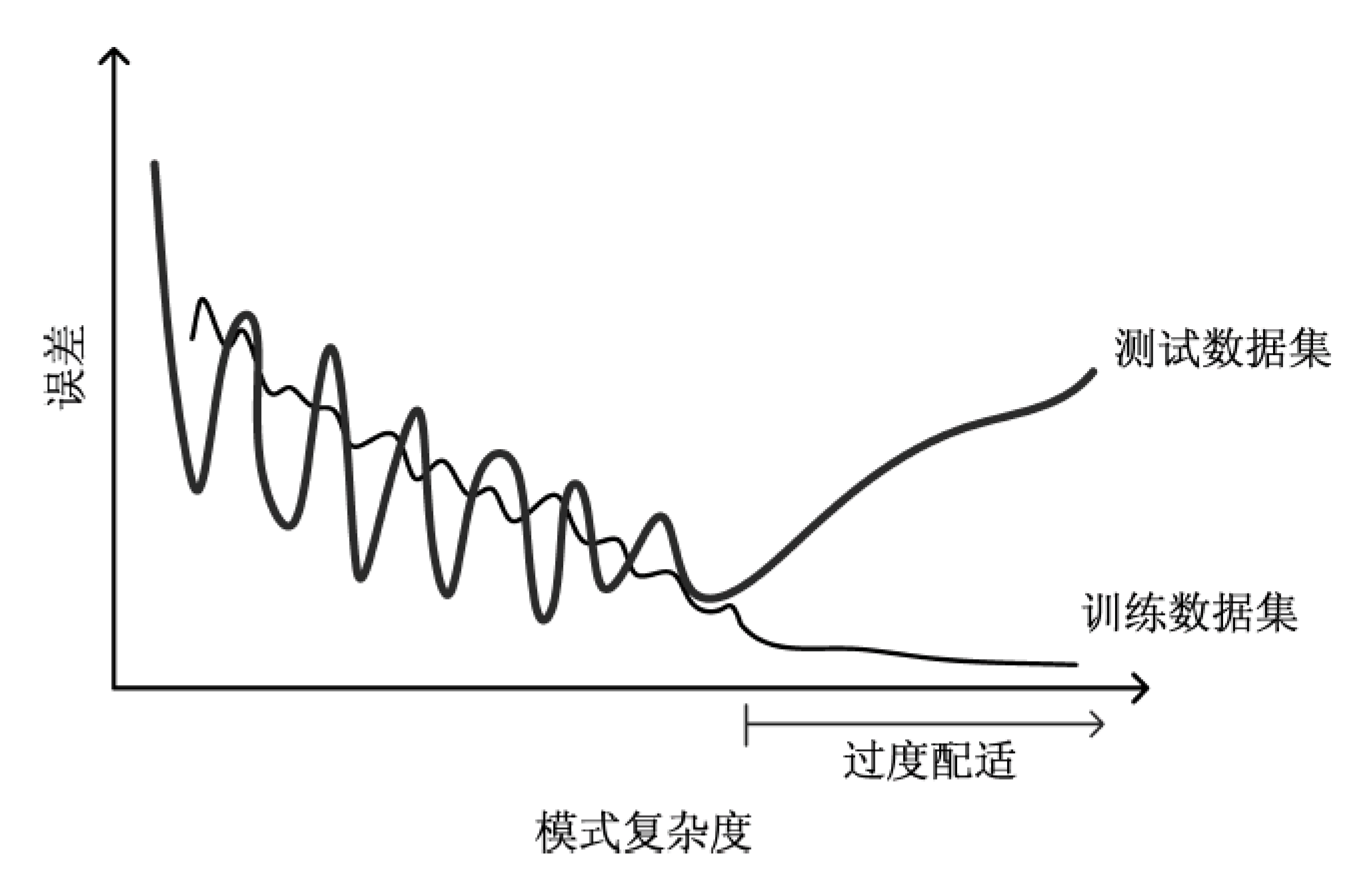


图 2.18 训练模型过度配适

数据分割的比例有不同的定义,均应代表原来的数据,一种方法是抽取 80%的数据用于建构模式,剩下的 20%则用于模式的效度检验。另一种方法为 k-fold 交互验证(k-fold cross-validation),如图 2.19。首先将数据分为 k 个等份,每次选取 $k-1$ 份数据进行模式训练,剩下的 1 份数据则用来测试模式,如此重复 k 次,使每笔数据都能成为训练数据集与测试数据集,最后的平均结果则用来代表模式的效度。这个方法的特例为当 k 个区间等于总样本数时,也就是每次选取 1 笔测试数据,称为"leave-one-out cross-validation",这个方法特别适用在样本个数很少的情况下,可有效涵盖整个数据,但缺点是计算时间长。

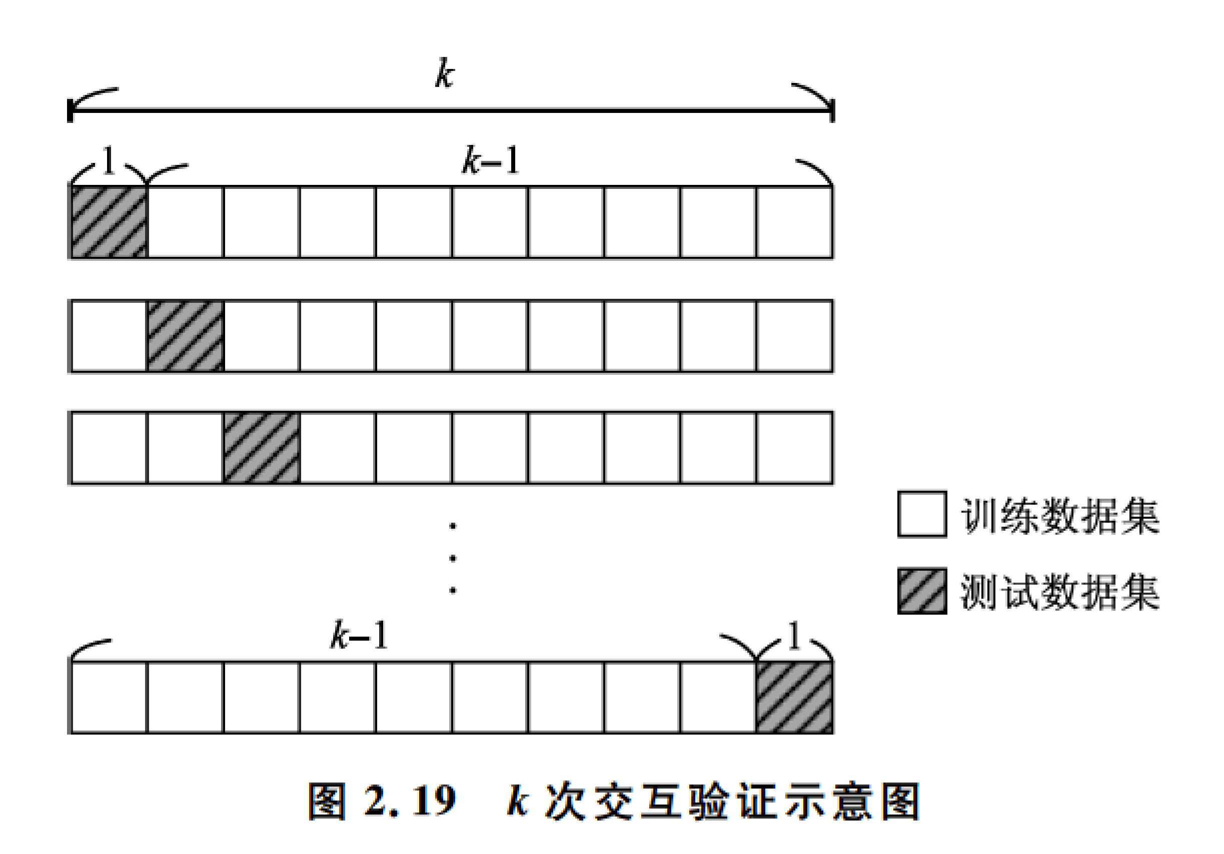


图 2.19 k 次交互验证示意图

2.11　应用实例——半导体厂制造技术员人力资源管理质量提升

2.11.1　案例背景

本案例以台湾省新竹科学园区某半导体公司的实际数据进行实证研究。该公司成立于1989 年，员工约有 3800 人，目前拥有一座六英寸①与两座八英寸晶圆厂，是全球非挥发性内存的主要供货商，提供从研发设计、制造生产到后端封装测试等一系列的完整服务。该公司制造部门的技术员来源复杂，有外籍劳工也有本地劳工，语言、文化、学历等背景皆不相同。有些主管常忙于制造现场的控制与管理，或处理较急迫的问题，而无暇兼顾人力资源的管理工作，甚至将技术员遴选的工作委托其他部门处理。其招募通常只用简单的英文、数学成绩与短短十几分钟的面试作为是否任用的依据。但面试者可能会有刻板印象或盲点，造成招募进来的员工素质参差不齐；若是使用事先规定好的问题照表操课，虽然可以降低面试者主观的因素，可是一则无法处理临时的情境反应，再则照本宣科对面试者没有自主权，较不易被主管接受。现场主管有时会抱怨新进技术员的素质无法符合公司的要求，希望能够招募适当的人员，以提升相关生产的绩效，却也无法具体提出技术员遴选的条件与方式。

本案例（简祯富等，2005）利用个案公司制造部门技术员的年龄、出生地、学历、科系、星座、血型以及之前的工作经验等个人基本数据与绩效数据，说明数据准备的实际应用过程。以生产线所有的技术员为对象，共计 465 位。数据搜集时间是从 2001 年 1 月 1 日至 4 月 3 日，数据源为该厂制造部人力训练组，数据属性说明于下：

（1）员工个人基本数据指针，包括：姓名、工号、课别（PHOTO、ETCH、DIFF）、班别（DA、DB、NA、NB）、职等（T1～T7）、国籍（本地劳工或外籍劳工）、生日、血型、毕业学校（school）、科系别（master）以及有无其他工作经验（experience）。

（2）工作表现与绩效指标，包括：提案次数（proposal）、特殊发现次数（apple）、操作错误次数（M. O.）、异常状况反映（report）以及绩效排名（ranking）等。

2.11.2　数据准备

1. 数据转换

转换数据格式以减少数据变化所产生的不必要的复杂度，转换方式如下。

（1）工作经验：工作经验原有数十种描述，例如，无经验、有某家半导体厂经验、纺织厂、会计等。将其简化成为三种类别，分别是“无”经验、“有”相关经验（有其他半导体厂经验）以及有“非相关”经验。

（2）学校：原有数十家，为简化数据与处理规则，分为专科、高职、高中三类。

（3）科系：原有 37 种科系，依据“教育部”的分类方式，分成艺术学类（FAA）、人文学类（H）、商业及管理学类（BA）、数学及计算机科学类（MCS）、医药卫生学类（MDT）、工程学类（E）、建筑及都市规划学类（ATP）、农林渔牧学类（AFF）、家政学类（HE）、运输通信学类（TC）、观光服务学类（ST）、大众传播学类（MC）与普通科（General）等。工程学类因所占比

① 1 英寸＝2.54cm。

例较大，再细分成工管类(IE)、电子工程类(EE)、化工类(ME)以及其他工程类(MO)，故共计有16类。

(4) 提案次数：提案次数从0～32次皆有，变异很大。进一步分类为：提案0次者以"never"表示，提案1～5次者以"seldom"表示，提案6～10次者以"sometimes"表示，提案11次及以上者以"often"表示。

(5) 异常状况反映：异常状况反映次数从0～11次皆有，变异也不小。再分类成：反映0次者以"never"表示，反映1～2次者以"seldom"表示，反映3～5次者以"sometimes"表示，反映6次及以上者以"often"表示。

(6) 特殊发现次数：特殊发现次数分布从0～4次，虽然变异不大，为增加其可读性，亦将之分类为：发现0次者以"never"表示，发现1～2次者以"seldom"表示，发现3次及以上者以"sometimes"表示。

2. 遗漏值的处理

本案例利用以下三种方式补值：

(1) 采推论的补值方式：目的在于以其他数据提供的信息，来估计遗漏值，并尝试以"较合理"的方式赋予补偿值意义。例如学校、科系数据不完整者，可检查其工号；其推论依据在于由于该部门技术员有许多是同校毕业生，且又一起报到并分发至同一班别者大多为相识的同学或朋友。因此也可用班别接近者的数据代入。

(2) 采平均值的补值方式：以平均值作为不偏估计量，让中心群数据来取代遗漏的数据。例如，考绩遗漏者以2或3代表。

(3) 不予补值的方式：例如血型的处理，在难以找到适当的处理方式时，可决定不予补值。

3. 数据特征强化

(1) 为了增加潜在有用的信息，生日部分以星座来表示，共12种星座；并进一步依其年龄区分：1956—1960年出生以"4B"表示，1961—1965年出生以"5A"表示，1966—1970年出生以"5B"表示，1971—1975年出生以"6A"表示，1976—1980年出生以"6B"表示，1981—1985年出生以"7A"表示，共六个年龄层。

(2) 在操作错误次数方面取得前两年的数据作补充，除了正式记载的操作错误数量外，生产线亦提供未经正式提报的数据。因此这一部分的数目远大于该部门去年至今操作错误次数的数目，有助于提升技术员操作质量分析的正确性。

(3) 技术员绩效部分，除了以年度考核为主的信息外，另参考非直属管理人员的意见，并分为四个等级：1、2、3、4，其中，1代表绩效最好，4代表绩效最差。未参加过绩效评比的新进人员，则以每月生产绩效表现之平均值为主，再加上直属主管的评核。本研究中，考绩等级为3者占所有样本的一半(50%)，考绩等级为2者则约有二成(21%)的人，考绩最好的1和最差的4则各有一成多的人，各绩效类别分布请参考图2.20。

整理过后的数据如表2.7所示。其中，前15项为输入属性，最后一项为输出属性。接着，利用数据可视化的方法，先对研究对象进行初步了解。在此465笔样本数据中，DIFF、ETCH与PHOTO部门各占1/3，日班与夜班的人数比例大致相当，本地劳工与外籍劳工的人数也是不相上下。职等则以中低职等的人为主，T1到T4职等占了近九成。至于

表 2.7　员工个人基本数据与工作绩效(部分数据)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
工号 TA ID	课别 Section	班别 Shift	职等 T. G.	国别 Nation	年龄 Age	星座 Star	血型 Blood	学历 School	科系类别 Master type	经验 Experience	提案次数 Proposal	特殊发现次数 Apple	操作错误次数 M. O.	异常状况反映 Report	考绩 Ranking
(略)	ETCH	DB	T7	Local	5A	AA	O	高职	商业及管理学类	非相关	Never	Never	0	Never	4
(略)	DIFF	DA	T6	Local	5A	GG	O	高职	商业及管理学类	非相关	Seldom	Never	0	Never	1
(略)	DIFF	DB	T6	Local	6A	EE	A	高职	商业及管理学类	非相关	Seldom	Never	0	Seldom	1
(略)	ETCH	DB	T6	Local	6A	GG	O	高职	商业及管理学类	非相关	Never	Seldom	0	Seldom	1
(略)	PHOTO	DB	T6	Local	5B	II	AB	高职	商业及管理学类	非相关	Never	Seldom	1	Seldom	2
(略)	PHOTO	DB	T6	Local	5B	EE	AB	高职	商业及管理学类	无	Seldom	Never	0	Seldom	1
(略)	DIFF	DA	T6	Local	5A	FF	A	高中	普通科	有	Never	Never	0	Seldom	3
(略)	DIFF	DA	T7	Foreign	5B	AA	O	高中	普通科	非相关	Seldom	Seldom	2	Seldom	1
(略)	DIFF	DB	T6	Local	5B	II	O	高职	商业及管理学类	非相关	Seldom	Never	0	Seldom	2
(略)	PHOTO	DB	T6	Local	5B	KK	B	高职	农林渔牧学类	非相关	Never	Seldom	0	Never	3
(略)	PHOTO	DB	T6	Local	5B	AA	A	高职	商业及管理学类	非相关	Never	Never	0	Never	4
(略)	DIFF	DB	T7	Local	5B	EE	A	高职	商业及管理学类	非相关	Never	Never	0	Never	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

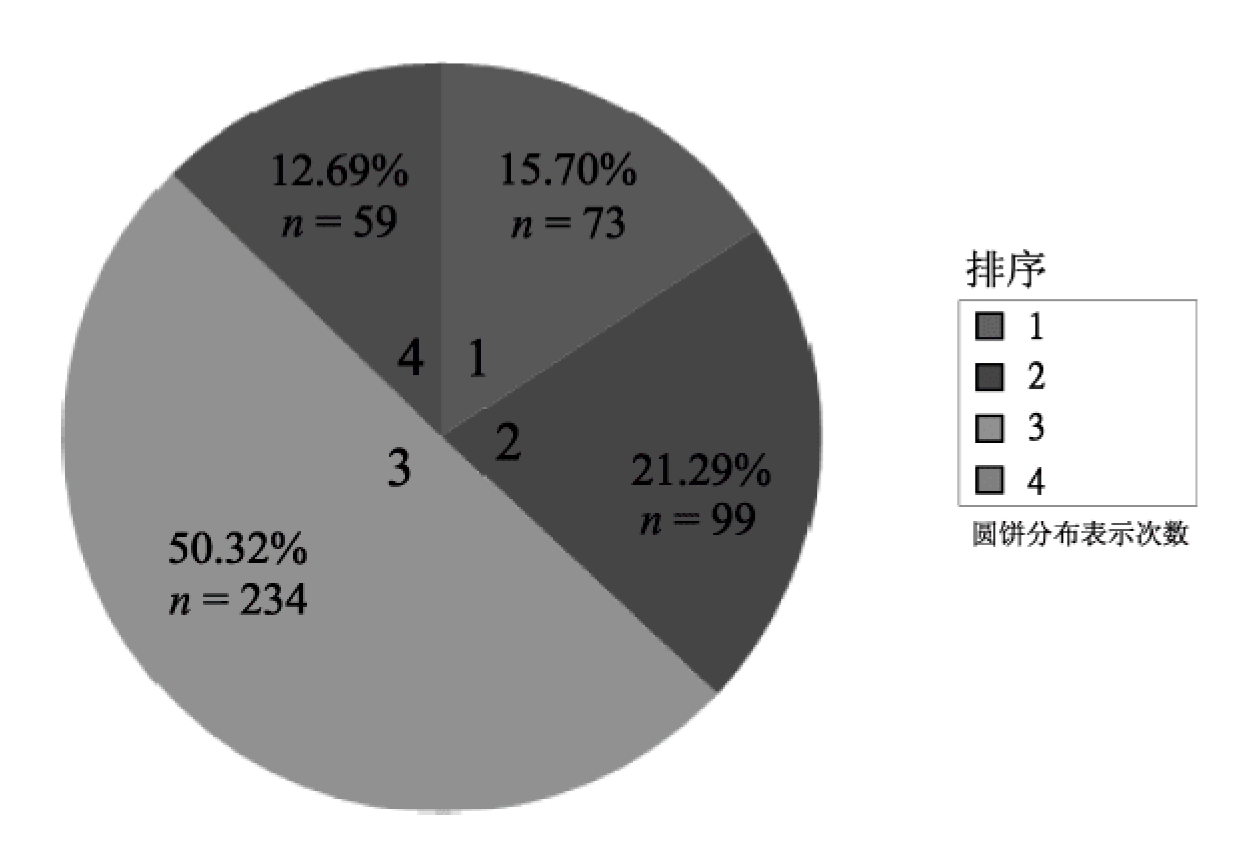


图 2.20　考绩等级分布

年龄层的分布以 1971—1985 年出生者居多,也就是现在年龄为 20～34 岁的人,约有 87%;血型以 O 型的人最多(36%),AB 型的人最少(6%);而星座则分布相当均匀,并无特殊集中现象。在学习经历背景方面,以高职(70%)与专科(18%)技职学校的学历为主,毕业科系以商管类(BA)较多,约占 1/3(32%),其次为电机类(MCS),约有 18%,其余则零散分布于各种科系类别中;而在所有技术员中,有非相关工作经验者约有一半(53%),其次为完全无经验者,约有四成(40%),只有非常少数的人是有相关工作经验的(7%)。

在工作表现方面,首先针对"提案次数"来看,约有 74%的人从未提案,22%的人属于"seldom",至于"sometimes"以及"often"的人相当地少;同样的现象也发生在"特殊发现次数"以及"异常状况反映次数"上,有近九成(88%)的人从来没有特殊发现,也有六成的人从来没有异常状况反映,在此两种指标上,"sometimes"以及"often"的人也同样相当地少。至于主管最关心的操作错误次数,从未操作错误的人约有八成(79%),有 16%的技术员曾经有过一次操作错误的记录,操作错误超过二次的人只占了相当少的比例(5%)。依据数据整理后的数据,即可进行后续的数据挖掘与模式构建。

2.12　结论

数据挖掘应从了解"数据"开始! 由于数据源的不同,数据挖掘分析时需处理的数据形态也不尽相同,例如离散型数据、连续型数据以及时间序列数据,随着数据挖掘技术的进步,也发展出不同形态数据的分析方法,适当地了解搜集的数据特性将有助于数据挖掘模式的选择,例如数据的维度多寡、数据分布、数据的变异程度等,而借由数据检查的步骤将可帮助数据分析人员采用最适当的分析工具,例如线性相关、长期趋势或周期循环等。随着越来越多的巨量数据产生,有意义的数据呈现已成为数据挖掘与巨量数据分析的重点,因此除了发展进阶数据挖掘分析方法外,可视化的工具将可提供数据挖掘分析者更多元的整合信息。

当明确定义问题与决定目标之后,必须对原始数据进行数据准备,转成数据挖掘工具可处理的形态,以改善数据质量,并使后续的分析工作更有效率。数据准备为数据挖掘的重要步骤,所需耗费的时间可能远高于其他步骤。此阶段所做的数据处理包括数据取得、检查数据与了解数据的形态与分布情形。数据清理目的在处理遗漏值、降低噪声数据以及纠正数

据的不一致性等问题，不论使用何种数据预处理的技术，都须尽可能使数据内容的损失最小。在检查数据方面，可根据所定义的属性，利用基本的统计图形检查数据的分布，删除会影响模式分析的变量，并进一步提供数据合并、数据转换或数据重新编码的信息；在数据预处理方面，数据清理包含遗漏值、空白值及离群值的处理。其中，由于空白值与遗漏值所代表的意义并不相同，不论是删除该笔数据或以特殊方式补值，对于挖掘结果的解释皆有不同的影响。而离群值的处理往往需借助领域专家的协助，以辨析该离群值为珍贵的信息，或仅是误植则可直接删除。

另外，数据维度亦影响挖掘模型的建立，一般而言，高维度的数据计算复杂亦较费时，因此如何降低维度是一项重要的课题，常见如利用主成分分析来降低维度。然而有些数据格式的转换，例如加工时间中的日期、班次、工序等，反而会增加数据的维度。因此，挖掘者需要进一步判断与决定信息的保存与数据的处理效率间的权衡。再者，有些数据挖掘模型，只能分析特定的数据格式，像是数字或文字、日期、时间等，所以在数据预处理时，也需了解数据的格式转换与其所代表的相对意义，以符合模式分析工具的需求。

问题与讨论

1. 试着举出既有的知识经由数据搜集、组织并整理后，形成信息之后该如何呈现？并指出过程中数据、信息、知识的组成元素及其所扮演的角色。

2. 数据挖掘可以处理的数据是根据不同的衡量尺度而被记录下来的数据，请举例说明各种不同的衡量尺度，包括名目尺度、顺序尺度、间距尺度、比率尺度及绝对尺度等，并说明可能的数据格式转换方法。

3. 试判断下列指标隶属于何种衡量尺度。

(1) 公元纪元

(2) 顾客满意度

(3) 每月薪资

(4) 竞赛名次

4. 假设有五个属性，分别为 A、B、C、D、E，并以7种函数得到不同的衡量结果，如下表。试以名目尺度、顺序尺度、间距尺度、比率尺度等四种不同的衡量尺度的角度加以判断并说明 $W_1(\cdot)\sim W_6(\cdot)$ 的衡量方式与 $V(\cdot)$ 的相同之处。

属性 / 衡量尺度	A	B	C	D	E
$V(\cdot)$	1	2	3	4	5
$W_1(\cdot)$	10	20	30	40	50
$W_2(\cdot)$	10	11	12	13	14
$W_3(\cdot)$	8	13	45	6	7
$W_4(\cdot)$	1	4	9	16	25
$W_5(\cdot)$	−10	−8	−6	−4	−2
$W_6(\cdot)$	22	25	28	31	34

5. 在现实中受到人为疏忽、记录设备异常等影响往往会造成数据的偏误,甚至是遗漏,请试举出三种以上不同类型处理遗漏值的方法。

6. 噪声值与离群值一直是数据清理的重要议题,试分辨并回答以下问题:

(1) 噪声值与离群值哪一个在分析上较具有意义? 为什么?

(2) 离群值也可能是噪声值吗?

(3) 噪声值也可能是离群值吗?

(4) 在何种情况下称作噪声值? 在何种情况下称作离群值? 请举一数据例子说明。

7. 某公司设计了一份问卷调查消费者使用某产品的满意度,将选项分为:“非常满意、满意、稍不满意、没意见与不满意”五项。试问该设计与李克特量表五点选项设计:“非常满意、满意、没意见、不满意与非常不满意”有无可能得出不同结论?

8. 如何找到数据较佳的特征(feature)或较低的维度(dimension)? 衡量较佳特征的依据为何? 新产生的特征如何借由过滤、增加既有数据集合或是融合(merge)特征成为另一新的数据集合?

9. 试比较极小值—极大值归一化与标准化转换后的值其范围有何差异?

10. 承上题,不同的值对于数据分析的结果或工具的使用各有哪些影响?

11. 在变量维度缩减中可采用逐步向前选择法、逐步向后删除法或综合两种方法,请问不同的选择法对结果各有何差异? 其可能的影响为何?

12. 随着计算机硬件技术的进步,许多复杂的计算开始能借由计算机高速计算能力得到解答,可分析的数据量也越来越大。若你是一位数据分析科学家,面对数百万笔数据,你会如何进行数据分析的第一步?

第 2 篇

数据挖掘方法与实证

第 3 章

关 联 规 则

关联规则(**association rules**)主要是从庞大数据中提取出一系列变量或因子间的关系,以探索数据的变量或项目间隐含的关系。阿格拉沃尔等(Agrawal *et al*.,1993a,1993b)最早从庞大事务数据中,发掘商品间隐含的关联规则,以了解消费者的购买行为与产品销售关系。关联规则是通过规则的描述所察觉的关联,即"若 A,则 B";例如,"若是下雨天,则雨伞销售量会增加"的关联规则,在日常生活中很容易可以发现类似的逻辑关联。

然而,有些实证的关联规则,如顾客购买"尿布",则常会一并购买"啤酒",即"尿布⇒啤酒"的关联规则,并不易事先察觉。经由卖场事务数据的数据挖掘发掘"啤酒与尿布同时出现在周末的同一笔交易中"的有趣现象,经过进一步了解后发现,有婴儿的美国家庭通常周末不会出去玩,因此在周末采购时,一方面买婴儿尿布,大部分的人又会顺便采购几箱啤酒在家里喝。根据所挖掘的关联规则,在卖场调整商品的陈设位置或做搭配营销,把啤酒和尿布摆在一起后,两者的销售量双双增加了三成。此外,在当当网买书时,网页会根据过去的交易记录找到关联规则,了解购买过此书的人同时也曾买过哪些商品,以自动推荐相关书籍或其他商品作搭配营销。通过数据挖掘的关联规则和有效推荐,不但能成功促销滞销品、增加商品的销售量,扩大营收,还可以预测顾客未来的购买行为,作为开发产品和进货的决策依据。

3.1 关联规则的定义与说明

关联规则定义如下,令 $I=\{i_1,i_2,\cdots,i_m\}$ 是所有相异物品项目(item)的集合,记载了以交易为主的相关数据,称为事务数据库(transaction database),为主要分析目标。T 表示一笔交易(transaction)记录内的物品项目集,有专属的代号(identification),且 $T\subseteq I$。假若在集合 D 中,项目 X 与项目 Y 产生关联规则,表示当交易记录 T 包含项目 X 时,有很大机会将同时包含项目 Y,此规则(rule)可表示为 $X\Rightarrow Y$ (if X then Y)。其中,X 为前提项目集(antecedent item set),Y 为结果项目集(consequent item set),X 和 Y 皆为 I 的子集合(或元素),且 $X\cap Y=\varnothing$。关联规则算法用词定义见表 3.1。

关联规则又称为**购物篮分析**(**market-basket analysis**),分析这些事务数据如同在卖场观察每一位顾客购物篮里究竟买了什么产品,如图 3.1 所示,每一个购物篮代表一位顾客在某个时间点的采购行为和一项交易记录,而且每位顾客购买的产品种类和数量不尽相同,购物篮分析即是从这些看似相关却又不尽相同的交易记录中,找出潜在有用的关联规则,以了解消费者购买行为的特定趋势及惯性,进而应用于营销、研发、供应链管理等相关决策上。例

表 3.1 关联规则算法用词的定义

用 词	定 义	用 词	定 义
TID(transaction identification)	每一笔交易的代号	C_k	k 阶候选项目集集合
D(database)	事务数据库	L_k	k 阶高频项目集集合
I(itemset)	项目集	H_i	i-项目集的散列表
k-项目集	k 阶项目集,项目集中包含 k 个项目	F	k 阶符合最小支持度的候选项目集

如,卖场商品的配置、销售配货、购物动线安排、产品定价及促销与相关宣传广告等,不仅可提升顾客对于卖场的整体满意度,也可提升卖场的销售利润。



图 3.1 购物篮分析示意图

假设[范例 3.1]为一有代表性的交易记录,则由记录可以很快地找出顾客消费行为间的特别模式,例如 3 个买牛奶的顾客中,有 2 个人买面包。然而,在大型卖场或网络拍卖中的交易记录时常达数十万笔以上,交易品项也达数百种。

[**范例 3.1**] 某大卖场 5 位客户的购买交易记录

交易记录	商品(代码)
101	牛奶(A)、面包(B)、饼干(C)、柳橙汁(D)
102	面包(B)、饼干(C)、汽水(E)、泡面(F)
103	牛奶(A)、饼干(C)、水果(G)
104	牛奶(A)、面包(B)、柳橙汁(D)、泡面(F)、水果(G)
105	饼干(C)、汽水(E)、水果(G)

[范例 3.1]中大卖场的交易可以整理为表 3.2 的二元数据表,每一列代表一笔交易,每一栏表示一个项目,若该项目出现在此交易中,则表示为 1,若没有则表示为 0。其中有 5 位顾客,共有 7 种商品,分别给予个别的编号代码。以事务数据库为宇集合,可求得此五位客户购买各商品的概率为 $P(A)=3/5$、$P(B)=3/5$、$P(C)=4/5$、$P(D)=2/5$、$P(E)=2/5$、$P(F)=2/5$、$P(G)=3/5$。特别注意的是,在关联规则分析中,重视的是消费商品之间的关联性规则,因此所感兴趣的为“商品项目”而非商品个数,故商品项目占总商品个数的比率于关联规则分析中为次要信息。

表 3.2　购物篮数据的二元数据表

交易记录	牛奶(A)	面包(B)	饼干(C)	柳橙汁(D)	汽水(E)	泡面(F)	水果(G)
101	1	1	1	1	0	0	0
102	0	1	1	0	1	1	0
103	1	0	1	0	0	0	1
104	1	1	0	1	0	1	1
105	0	0	1	0	1	0	1

3.2　关联规则的衡量指针

关联规则常利用支持度、置信度和增益等三个衡量指标来分别表示其显著性(significance)、正确性及价值，通过给定**最小支持度(minimum support)**与**最小置信度(minimum confidence)**作为支持度与置信度的门槛值(minimum threshold)，再评估该规则的信息价值和增益。若该规则的支持度与置信度大于或等于分析人员所规定的门槛值，表示该规则有助于进行推论，若该规则的增益满足大于 1 的条件，则表示其发生的条件概率有比原先的概率提高，亦即该规则有效。关联规则的分析可以提供一序列或矩阵关系的品项相关矩阵，让决策者了解品项间的关联关系，以营销策略或卖场配置方案，提升获利及客户满意度。关联规则三项衡量指针的计算公式与物理意义阐述如下。

(1) **支持度(support)**：支持度衡量前提项目 X 与结果项目 Y 一起出现的概率 $P(X\cap Y)$，表示该规则在全部交易记录中出现的比率，如式(3.1)所示。支持度表示关联规则相对于全部数据必须具有一定的普遍性(即具显著性)，才是有效的信息。最小支持度门槛主要用于管控关联规则所必须涵盖的最少数据比率；其可删除所占比率偏低的关联性，以撷取出较具代表性的关联规则于实务应用。

$$Support(X \Rightarrow Y) = P(X \cap Y) \tag{3.1}$$

在表 3.2 的交易记录中，若欲了解消费者购买牛奶(A)的同时也会选购面包(B)的规则是否具有显著性，可通过支持度衡量值，即计算顾客同时购买牛奶与面包的概率，计算如下：

$$Support(\text{牛奶} \Rightarrow \text{面包}) = P(\text{面包} \cap \text{牛奶}) = \frac{2}{5} = 0.4$$

(2) **置信度(confidence)**：置信度衡量前提项目 X 发生的情况下，结果项目 Y 发生的条件概率，即 $P(Y|X)$，表示对当前提项目 X 发生时，可推得结果项目 Y 的规则的正确性的信心程度，如式(3.2)所示。置信度是衡量关联规则是否具有可信度的指标；因此，置信度须达到一定水平(通常为 0.5)，利用最小置信度为门槛去除正确概率较低的关联规则。

$$Confidence(X \Rightarrow Y) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \tag{3.2}$$

在表 3.2 的交易记录中，若欲了解规则“消费者购买牛奶(A)后也会选购面包(B)”的信心程度，可依式(3.2)衡量其置信度，衡量结果表示在消费者购买牛奶的情况下，也会购买面包的概率为 0.667。

$$Confidence(牛奶 \Rightarrow 面包) = P(面包 \mid 牛奶) = \frac{2/5}{3/5} = 0.667$$

(3) **增益(lift)**：增益衡量用于比较置信度与结果项目 Y 单独发生时两者概率间的大小，即 $P(Y|X)/P(Y)$，如式(3.3)所示。增益值的物理意义是比较关联规则置信度与原本结果项目 Y 发生的概率以衡量该规则的价值和相对效益，因此增益值至少要大于1，表示该关联规则的预测结果比原本表现好，亦即其置信度大于原本结果项目 Y 发生的概率(Berry & Linoff，1997)。

$$Lift(X \Rightarrow Y) = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{3.3}$$

在表3.2的交易记录中，消费者购买牛奶(A)后也会选购面包(B)的规则的增益为1.111，计算如下：

$$Lift(牛奶 \Rightarrow 面包) = \frac{P(面包 \mid 牛奶)}{P(面包)} = \frac{2/3}{3/5} = 1.111$$

进行关联规则挖掘时，通常会先设定挖掘所得的规则的支持度与置信度的门槛值，以作为挑选关联规则的准则。由此筛选出的规则必满足决策者规定的最小支持度和最小置信度。当满足这两个条件后，再判断这些规则的增益值是否大于1；大于1则保留，反之删除。当三个指标皆成立时，即为所推导的关联规则。在此例中，若分析人员设定支持度与置信度的门槛值为0.2与0.5，则此规则“顾客于购买牛奶的同时也会选购面包”将被列为显著信息，置于有效信息的集合中。而规则的增益值为1.111>1，经过最终衡量后，此规则将被列为显著信息，置于有效的信息集合中。

由[范例3.1]可以进一步了解更多商品的关联规则。例如，牛奶(A)的支持度为0.6、面包(B)的支持度为0.6、饼干(C)的支持度为0.8、牛奶和面包($A \cap B$)的支持度为0.4、牛奶和饼干($A \cap C$)的支持度为0.4、面包和饼干($B \cap C$)的支持度为0.4以及牛奶和面包和饼干($A \cap B \cap C$)的支持度为0.2，则可推导出衡量“若牛奶与面包则饼干”的关联规则的三项指针值如下：

$$Support(牛奶,面包 \Rightarrow 饼干) = P(牛奶,面包,饼干) = 0.2$$

$$Confidence(牛奶,面包 \Rightarrow 饼干) = P(饼干 \mid 牛奶,面包) = 0.5$$

$$Lift(牛奶,面包 \Rightarrow 饼干) = \frac{P(饼干 \mid 牛奶,面包)}{P(饼干)} = \frac{0.5}{0.8} = 0.625$$

由于增益值为0.625<1，此规则“顾客于购买牛奶的同时也会选购饼干”。在经过最终衡量后，将被列为不显著信息，排除于有效的信息集合中。

其他规则的支持度、置信度以及增益值都可利用同样的计算方式求得。表3.3列出四项管理者有兴趣了解的规则的测量情况；其中，可看出所有包含三项商品的规则都没有显著增益效果，从此事务数据库中唯一提取得到的关联规则仅包含两项商品，即“若消费者购买牛奶，则也会购买面包”。由于此规则的增益值为1.111，可知在消费者已经购买牛奶的情况下，购买面包的概率会是原本的1.111倍。通常在欲探讨的关联规则中，商品项目越少时，消费该商品组合的顾客人次会相对提升，该规则的显著性会越强烈。

表 3.3　四条规则的增益测量

规　　则	支持度	置信度	增益
若牛奶(A)与面包(B)则饼干(C)	20%	50%	0.625
若牛奶(A)与饼干(C)则面包(B)	20%	50%	0.83
若面包(B)与饼干(C)则牛奶(A)	20%	50%	0.83
若牛奶(A)则面包(B)	40%	67%	1.111

关联规则分析广泛应用在零售业与大型卖场的数据挖掘，借由分析后所取得的信息，得知顾客所偏好的产品与其他产品间的关联，以制订良好的市场营销及配售计划。

3.3　关联规则的类型

关联规则可以分成三种类型(Han & Kamber,2006)，分述如下。

1. 以规则中属性值的形态为基础

布尔关联规则(Boolean association rule)系指关联规则中的数据集合属性皆为布尔值，仅探讨“项目是否出现”，如0或1。[范例3.1]的关联规则，均为由购物篮分析所得的布尔关联规则，如牛奶⇒面包(支持度为40%，置信度为67%，增益值为1.111)，并未区分消费者所购买的牛奶和面包的数量与价值。

若所要描述的规则为属性值的关联性或项目在数量范围下所产生的相关性，则称为**量化的关联规则(quantitative association rule)**。在布尔关联规则中，可以视需要将项目或属性的值分为数个不同子项目；例如牛奶与面包可以根据不同厂牌编为不同群组，以建立较细的关联规则，如“牛奶 $A1$ ⇒面包 $B1$”、“牛奶 $A2$ ⇒面包 $B2$”。

表3.4为卖场顾客基本数据中年龄的量化属性值，通过数值化的区间划分与归类后，此量化属性可转换为布尔属性(Agrawal & Srikant,1996)。然而，经过布尔值转换所挖掘出的关联规则无法看出消费者真实年龄，仅能看出其年龄区间。

表 3.4　消费者年龄与转换后的布尔属性值

编号	消费者年龄	年龄[10,20)	年龄[20,30)	年龄[30,40)	年龄[40,50)
1	18	1	0	0	0
2	35	0	0	1	0
3	26	0	1	0	0
4	46	0	0	0	1

2. 以规则中所涵盖的数据维度为基础

根据关联规则所涵盖的数据维度来分类，若规则的项目或属性针对单一维度时，称之为**单一维度关联规则(single dimensional association rule)**。例如，“购买牛奶与面包⇒购买饼干与巧克力”，只有着眼于“购买”此一维度。反之，若关联规则中的项目或属性着眼于两个或两个以上维度时，则称为**多维度关联规则(multidimensional association rule)**或**复合维度关**

联规则。例如,客户的信息包括身份、性别、收入、所购物品等,可将记录中的每个属性或维度看作一个规则的依据。例如,银行的理财专员应用数据挖掘发现:"单身、三十岁以上的工程师,年收入30万至40万人民币之间⇒购买海外基金"的比例特别高,即为一多维度关联规则。

3. 以规则集合中所涵盖的抽象层级为基础

若规则属性或项目隶属于同一层级,称为**单一层级关联规则**(**single-level association rule**);例如,购买牛奶⇒购买面包,可从中得到较具体与精确的信息。在实务上,数据可能同时包含较低阶层和较高阶层的项目集集合,称为多阶层数据。针对多阶层数据,分析者可先建立**概念层级树**(**concept hierarchy tree**),作为挖掘规则的架构。由于所搜集的数据未必充足,当数据隶属于较低阶层的项目集集合不易发现关联规则时,即可借由提升交易项目的层级,以发现较明显的关联规则。以流行用品拍卖场事务数据分析为例,可先推导概念层级树如图3.2所示,再建立**多阶层关联规则**(**multilevel association rule**)。

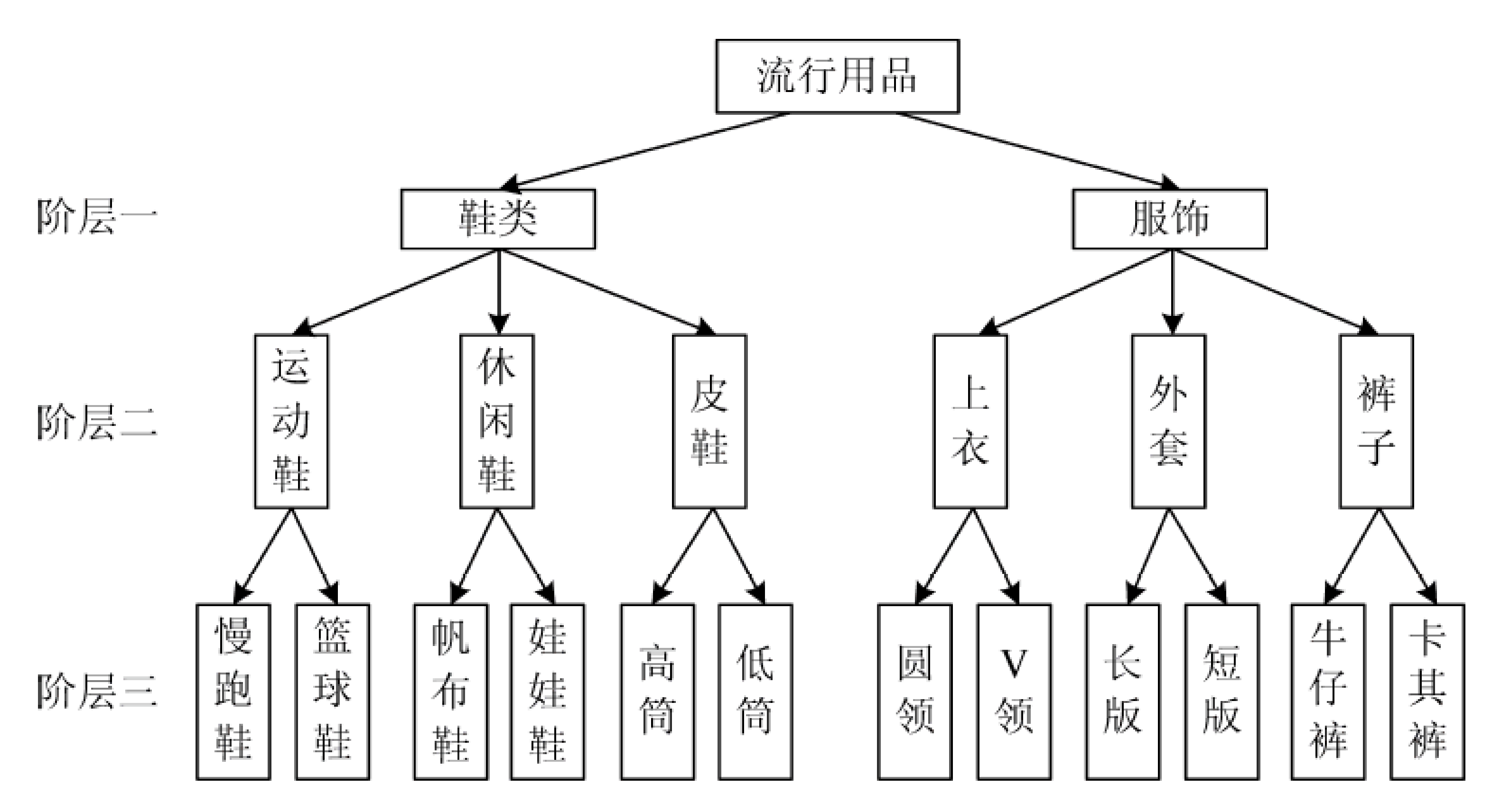


图3.2 流行用品拍卖场事务数据的概念层级树

3.4 关联规则算法

关联规则是从搜索的可能规则中,根据其支持度、置信度和增益等衡量指标,筛选出具有足够支持度的所有**高频项目集**(**frequent itemsets**),从中找出属性或项目间有所关联的规则。

为了避免产生的规则过于繁多导致无法凸显真正重要的规则,必须适当地定义最小支持度以过滤多数次要的规则,同时产生的规则的置信度与增益值必须高于决策者给定的最低门槛值,规则才能成立。定义出相关门槛值之后,即可据此搜索数据库中符合条件的关联规则。

关联规则算法主要由搜索方式、计算项目及支持度来组成,好的搜索算法可有效率地构建出有用的关联规则。说明如下。

1. 搜索方式

数据越庞大,需要搜索的属性或项目组合也相对复杂,如何进行搜索为影响关联规则的

建立结果的重要关键。搜索方式主要分为**广度优先搜索**与**深度优先搜索**两大类。广度优先搜索方式为由下往上搜索，此种搜索方式在计算 k 个项目集的支持度前，必须先算出 $k-1$ 项目集的支持度，才能由下往上找出项目集之关联规则；深度优先搜索方式为由上往下搜索，此种搜索是以递归的方式顺着所构建的树状数据结构，由上而下寻找并计算项目集的支持度，以找出显著的关联规则。

2. 计算项目及支持度的方式

计算项目及支持度的方式分为**水平数据配置**与**垂直数据配置**两大类。水平数据配置方式是以计算项目发生次数来提升算法效率。将计数器的初始值设为0，之后扫描所有事务数据，假若某笔交易内存在显著的项目时，该计数器的数值即从0开始往上累加；垂直数据配置方式是以交集找出显著项目集所组成的关联规则，皆以升幂的方式储存其事务数据代号，以提升整体效率。

除了阿格拉沃尔等(Agrawal *et al.*，1993a，1993b)最早提出的 Apriori 算法外，已发展其他各种关联规则算法，例如，Partition 算法(Savasere *et al.*，1995)、DHP(direct hashing and pruning)算法(Park *et al.*，1995)、MSApriori 算法(Liu *et al.*，1999)以及 FP-Growth (Han *et al.*，2000)等。其中，Partition、DHP 及 MSApriori 算法都是以 Apriori 算法为基础所发展的广度优先搜索算法，其搜索方式均为由下往上搜索高频项目集以及候选项目集(candidate itemset)，以找出显著的关联法则，整理如表3.5。

表3.5 关联规则算法与特性

算 法	作者(年代)	主要特色	搜索方式	数据配置方式	缺点或限制
Apriori	Agrawal *et al.* (1993a，1993b)	反复产生候选项目集，找出所有高频项目集，进而推导规则	广度优先	水平数据配置	需反复搜索数据库，花费I/O时间
Partition	Savasere *et al.* (1995)	将数据库分区段，找出各区段的高频项目集加以集合，再次搜索数据库找出真正高频项目集	广度优先	垂直数据配置	在各区段中会产生较多的非相关项目集
DHP	Park *et al.* (1995)	利用散列表(hash table)删减不必要的候选项目集	广度优先	水平数据配置	一开始需花时间建立散列表
MSApriori	Liu *et al.* (1999)	在数据项出现频率不一致的情况下，挖掘低频率但重要事件之关联规则	广度优先	水平数据配置	需多加探讨多重最小支持度与算法中参数的主观制定
FP-Growth	Han *et al.* (2000)	频率样式成长为算法的演绎基础，可改善Apriori无法有效处理大量数据的缺点	深度优先	水平数据配置	挖掘过程中需较多的额外处理时间及储存空间来存放FP-tree

3.4.1 Apriori 算法

Apriori 算法(Agrawal & Srikant,1994)为挖掘高频项目集的布尔值关联规则中最具代表性的算法,随后发展的关联规则算法大多以其为基础。Apriori 算法的主要概念是在大量的数据集中,利用项目集来建立关联规则,并计算每一个候选项目出现的数目,依据所设定的最小支持度为门槛,来衡量候选项目的关联规则是否显著。

随着数据项的不同,可定义的项目集也会有所不同。当项目个数越多,产生的项目集合数量也会越庞大,若逐一计算所有有兴趣的项目集的支持度将非常缺乏效率。因此,Apriori 算法采用水平方向进行项目集的搜索(level-wise search);其方式是通过 k 项目集(k-itemset)的组合去探索 $k+1$ 项目集,以提升发现高频项目集的效率。Apriori 算法由单一项目集(1-itemset)开始,反复产生候选项目集与搜集项目集的步骤,直到找出所有高频项目集为止,即无法找到更高频的显著项目集时。首先以联合(join)的方式产生候选项目集,候选项目集的支持度必须大于或等于用户所定的最小支持度,例如包含项目 E 的数据个数占总数据个数的比例大于最小支持度,则称项目集$\{E\}$为高频项目集,如式(3.4)所示:

$$\frac{|T(E)|}{|T|} \geqslant \text{min Support} \tag{3.4}$$

其中,$|T(E)|$代表数据中包含 E 的个数;$|T|$表示数据集中的总事件个数。同样地,如果同时包含项目 E 和 F 的数据个数占总数据个数的比例大于最小支持度时,则事件集$\{E,F\}$为高频项目集。由于事件集$\{E\}$与$\{E,F\}$皆为高频项目集,可就其所包含的项目个数来区分;如区分$\{E\}$为高频 1-项目集;而$\{E,F\}$则为高频 2-项目集。同理,若某高频项目集里包含 k 个数据项,则称为高频 k-项目集。

为了改善产生高频项目集的效率,Apriori 算法应用类似递移律的概念,称为**反单调性**:若某候选项目集为高频,则其所有的子集合必定是高频项目集。也就是若$\{E,F\}$为一高频项目集,则$\{E,F\}$内的任一非空子集也会满足高频项目集的特性;反之,若某项目集之任一子项目集为非高频项目集,则该项目集亦为非高频项目集。即若$\{E\}$或$\{F\}$有任一项目集为非高频项目集,则$\{E,F\}$也必为非高频项目集。根据此特性可对候选项目集进行进一步检查或删除以产生高频项目集。最后,从高频项目集中即可产生一系列的规则,若这些规则满足所规定的最小置信度与增益值大于 1,则视为有效的关联规则。

Apriori 算法建立关联规则主要可分为五个步骤,如图 3.3 所示。

(1) 快速地扫描事务数据库,找出所有 1-项目集后(此处需注意 Apriori 算法是采用由下往上的方式搜索项目集,故其第一个项目集常仅包含单一商品),再与所规定的最小支持度作比较,若通过门槛则可视为高频项目集,又称为高频 1-项目集,记为 L_1。设定 $k = 1$。

(2) 设定 $k=k+1$ 并产生新的候选 k-项目集;删除候选 k-项目集内有任意$(k-1)$-子项目集不属于 L_1 的候选项目集,并记过滤完后的候选 k-项目集为 C_k。

(3) 计算 C_k 集合中的各自对应的支持度 S 是否大于或等于用户所定的最小支持度,以得到高频项目集集合,而搜集 C_k 内符合条件限制的项目集即称为高频 k-项目集,或记为 L_k;C_k 内不符合条件限制的项目集则删除。

(4) 判断是否已搜索过所有的候选项目集。若已搜索完所有可能的候选集,则继续步骤(5);若否,则回到步骤(2)。

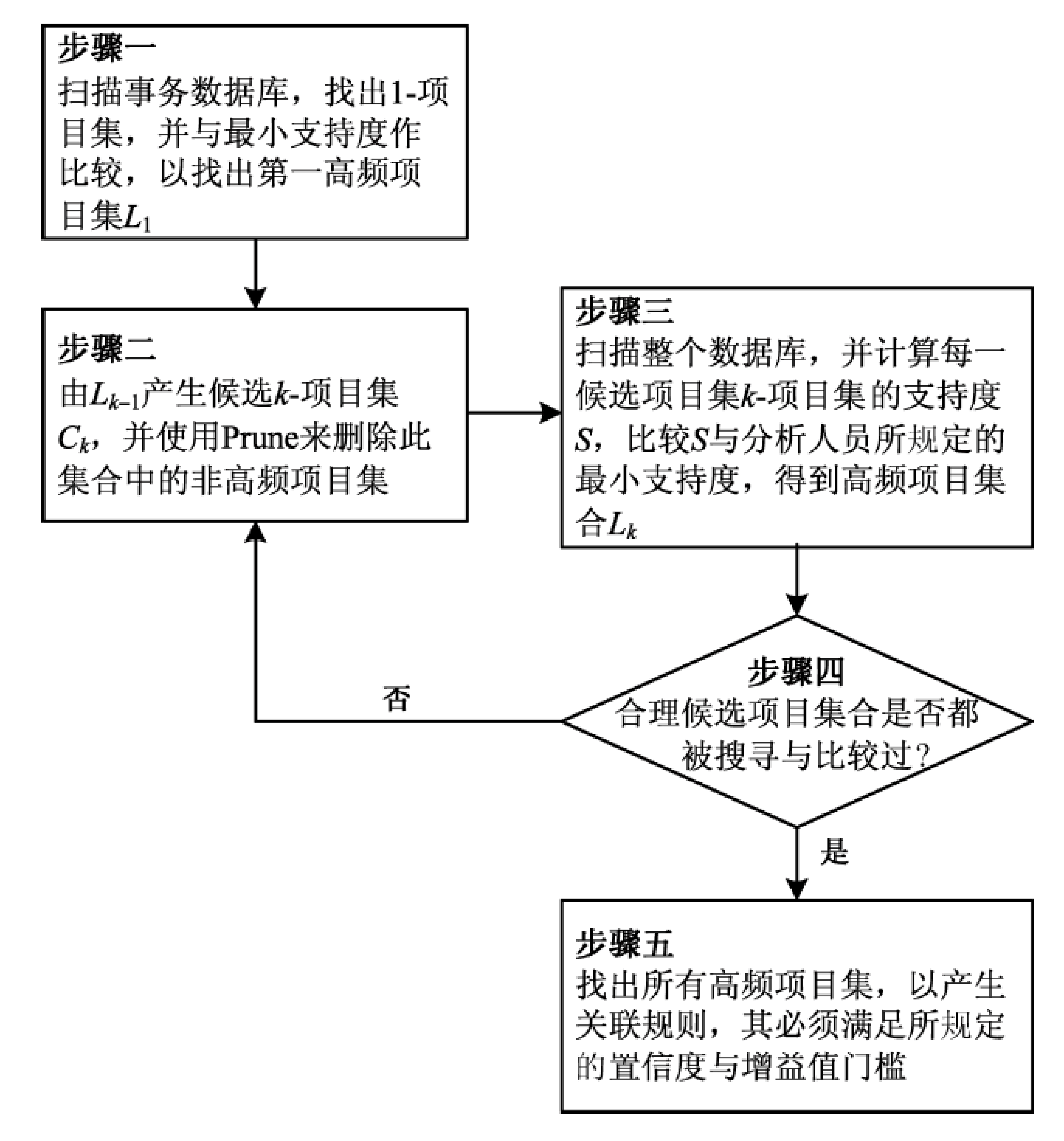


图 3.3　Apriori 算法流程图

(5) 计算所搜集的项目集的置信度与增益值，找出具显著性的关联规则以帮助管理者规定相关决策。

以某购物中心的交易记录为例，如表 3.6 所示，其中包含 4 笔交易记录与 5 种商品，每一品项专属的代码如括号所示。

表 3.6　购物中心交易记录

交易记录	商品(代码)
201	巧克力(A)、饼干(C)、汽水(D)
202	牛奶(B)、饼干(C)、面包(E)
203	巧克力(A)、牛奶(B)、饼干(C)、面包(E)
204	牛奶(B)、面包(E)

首先，在与购物中心的决策者讨论后，定义出最小支持度为 0.5，且最小置信度为 0.5。接着利用 Apriori 算法对此购物中心消费记录，进行高频项目集集合的搜索与删除，如图 3.4 所示。

(1) 首先将交易记录转换成代码或布尔值表示的离散数据，如图 3.4(a)。再以由下往上搜索的方式，从基层的单项商品组合开始建立 1-项目集的集合，可得 C_1 并计算出各项目集所对应的支持度，如图 3.4(b)所示。接下来比较所得支持度与所定支持度门槛 S 来决定高频项目集。从图 3.4(c)可看出经过搜索后，可以得到高频 1-项目集有$\{A\}$、$\{B\}$、$\{C\}$、$\{E\}$，将其记为 L_1。

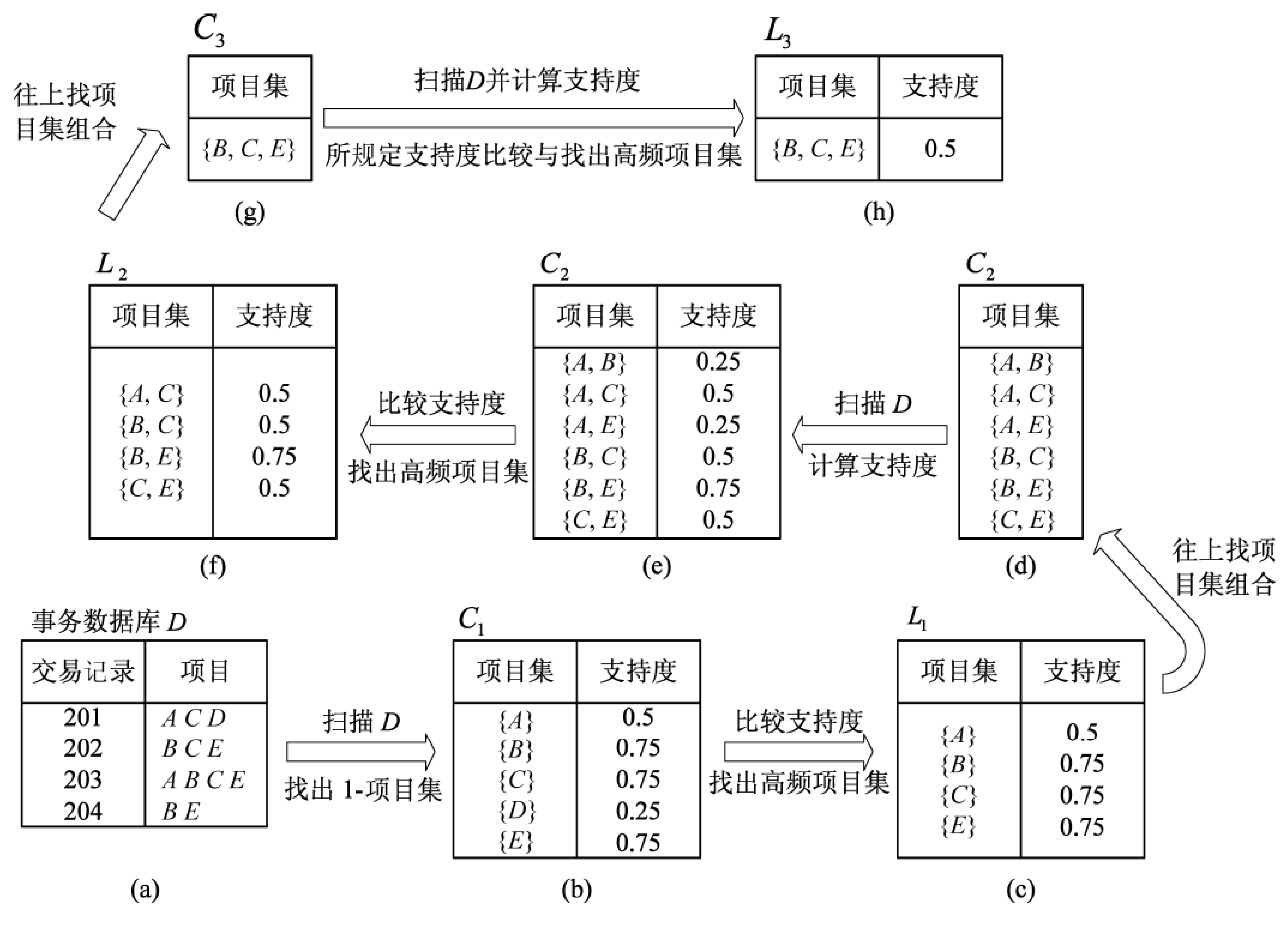


图 3.4 候选项目集集合和高频项目集集合的产生过程

(2) 之后往上推一层，将所得的高频 1-项目集组合成 6 个 2-项目集(2-itemset)，如图 3.4(d)所示，记为 C_2；接着计算其支持度，如图 3.4(e)所示。得到第二层各项目集所对应的支持度后，与支持度门槛 S 比较以决定高频 2-项目集，如图 3.4(f)所示，并记为 L_2。

(3) 继续往上搜索，确认包含三个项目的项目集是否也会符合高频项目集的特性。再通过 L_2 中各项目集往上搜索后，发现仅能找到一个第三层的项目集，即为{B,C,E}，记为 C_3，如图 3.4(g)所示。在此，不须将项目集{A,C,E}列于 C_3 中，因为其子项目集{A,E}并非高频项目集，因此可事先删除其成为高频项目集的可能性。由于项目集{B,C,E}的子项目集{B,C}、{B,E}以及{C,E}皆为高频项目集，因此{B,C,E}亦有机会成为高频项目集。最后，通过计算项目集{B,C,E}的支持度 S，与最小支持度比较后可得最上层的高频 3-项目集集合为{B,C,E}，并记为 L_3，如图 3.4(h)所示。

(4) 接下来利用所找到的高频 3-项目集{B,C,E}来建立关联规则。在此例中，共有 12 种可能的规则，依序计算这些规则所对应的置信度与增益值，如表 3.7 所示，可从中找出 6 条显著的关联规则。

表 3.7 12 条规则的支持度、置信度与增益值量测值

规　　则	支持度	置信度	增益值
若牛奶(B)则饼干(C)	0.5	0.667	0.889
若牛奶(B)则面包(E)	0.75	1	1.333

续表

规　　则	支持度	置信度	增益值
若饼干(*C*)则牛奶(*B*)	0.5	0.667	0.889
若饼干(*C*)则面包(*E*)	0.5	0.667	0.889
若面包(*E*)则牛奶(*B*)	0.75	1	1.333
若面包(*E*)则饼干(*C*)	0.5	0.667	0.889
若牛奶(*B*)则饼干(*C*)与面包(*E*)	0.5	0.667	1.333
若饼干(*C*)则牛奶(*B*)与面包(*E*)	0.5	0.667	0.889
若面包(*E*)则牛奶(*B*)与饼干(*C*)	0.5	0.667	1.333
若牛奶(*B*)与饼干(*C*)则面包(*E*)	0.5	1	1.333
若牛奶(*B*)与面包(*E*)则饼干(*C*)	0.5	0.667	0.889
若饼干(*C*)与面包(*E*)则牛奶(*B*)	0.5	1	1.333

Apriori 算法采用水平的广度搜索法，以逐层扩展的方式来搜索高频项目集；并利用反单调性原理进行较完整的候选项目集的删减。然而，其主要缺点在于逐层扩展候选项目集必须大量重复地搜索数据库，因此当高频项目集长度较长或数据量较多时，即必须花费较长的时间来挑选产生候选项目集。因此，许多算法即以改善此缺点发展而来，以下介绍四种较具代表性的改良算法。

3.4.2　Partition 算法

为解决 Apriori 算法直接对整体数据进行高频项目集的搜索因而效率不彰的问题，分析者可以通过适当的方式将数据分为若干小群，再从这些小群中分别搜索高频相关群，最后再将这些从小群所搜索的高频相关群合并并加以评估即可得到所要的结果。Partition 算法(Savasere *et al.*，1995)以多次小群的搜索过程取代并降低整体数据的搜索过程，可有效减少计算时间。若 X 为数据库 D 的一高频项目集，当 D 被切为数个“分割”$P_1, P_2, \cdots, P_n$ 后，则 X 至少为一个分割 P_i 的高频项目集，Partition 算法先进行数据分割，再进行扫描与找出高频项目集，以建立显著关联规则。

Partition 算法将数据库 D 分割为许多区段，容纳于主存储器中，再于内存中一次处理一个分割，主要包含以下两阶段：

(1) 将数据库分成多个互不相交的时间区段，并分别计算区段中相关项目集的支持度，以找出各区段中的高频项目集，称为**区域高频项目集(local frequent itemset)**；其中，分割区块的大小与个数取决于计算机内存大小。第一次对事务数据库进行扫描时，此算法的主要工作是读取每一个分割 P_i，并逐层搜索找出该分割中的区域高频项目集集合，记为 L_{P_i}。以表 3.8 的复合式餐饮店数据为例，此算法的第一阶段为以顾客及时间区段，将原始事务数据形态转换成 P_1、P_2、P_3 的分割时段数据，并计算每一个时段中相关项目集的支持度。如于时段 P_1、P_2、P_3 中，分别可找出的高频项目集为 L_{P_1} ={三明治，豆浆}、L_{P_2} ={蛋糕，咖啡}、L_{P_3} ={面包，奶茶}。

表 3.8　复合式餐饮店分割时段的数据库形态

顾客	交易情况	时　段
A	三明治、豆浆	P_1(5:00～7:00)
B	三明治、奶茶	P_1(5:00～7:00)
C	汉堡、咖啡	P_1(5:00～7:00)
D	蛋饼、奶茶	P_1(5:00～7:00)
E	三明治、蛋饼、豆浆	P_1(5:00～7:00)
F	蛋糕、咖啡	P_2(7:00～9:00)
G	汉堡、面包、咖啡	P_2(7:00～9:00)
H	三明治、奶茶	P_2(7:00～9:00)
I	面包、蛋糕、咖啡	P_2(7:00～9:00)
J	蛋饼、蛋糕、咖啡	P_2(7:00～9:00)
K	汉堡、蛋饼、奶茶	P_3(9:00～11:00)
L	蛋饼、豆浆	P_3(9:00～11:00)
M	面包、三明治、奶茶	P_3(9:00～11:00)
N	汉堡、蛋糕、咖啡	P_3(9:00～11:00)
O	面包、奶茶	P_3(9:00～11:00)

(2) 取所有区域高频项目集的并集，即{$L_{P_1} \cup L_{P_2} \cup \cdots \cup L_{P_n}$}，以产生 D 的整体候选项目集集合。对 D 重新计算各候选项目集的支持度，以搜索数据库的真正的高频项目集(global itemset)。如上例，可并集 L_{P_1}、L_{P_2} 与 L_{P_3} 以得 D 中的整体候选项目集集合 L={{三明治，豆浆}，{蛋糕，咖啡}，{面包，奶茶}}，再经由 D 的整体数据对 L 内的候选项目集进行支持度评估，以确定这些项目集对于整体数据的支持度高于所设定的门槛。评估后可得仅{蛋糕，咖啡}在整体数据中为高频项目集，因此便可根据此结果评估置信度与增益度以找出显著的关联规则。

整体来说，Partition 算法最多仅需在事务数据库进行两次完整搜索即可找出所有区域高频项目集集合；若所有的分割所得的区域高频项目集集合均相同，则仅需完整扫描数据库一次即可。

Partition 算法与 Apriori 算法的概念极为相似，但应用“切割”的概念将事务数据分割成一些没有重叠的部分，使得主存储器运作时能加快速度，降低扫描整个数据库的次数，其优点是可大幅提升关联规则的搜索效能；但缺点为若在各区段中产生太多的非相关项目集时，则需要大量的储存空间。

3.4.3　DHP 算法

当事务数据库 D 中的交易记录很多时，Apriori 算法产生的候选 2-项目集及其他高阶项目集的数量将会非常庞大。同时，计算候选 k-项目集出现次数时需要搜索整个 D，因而需要花费相当高的处理成本。DHP(direct hash-based pruning)算法主要是以散列(hash)的

技术，减少记录候选 2-项目集所占用的空间、删除不必要的候选 2-项目集，以改善 Apriori 算法的搜索效率；相关的散列技术包含**散列树**（**hash tree**）以及**散列表**（**hash table**）（Park *et al.*，1995）。

若以某购物中心为例，表 3.9 为 2-项目集的散列表形式范例，分析者需先决定散列函数（hash function），假设选择除留余数作为散列函数为 $h(x,y)=[(x\ order)\times10+(y\ order)]\bmod 7$，其中，$x\ order$ 与 $y\ order$ 分别代表 2-项目集的顺序，以项目集$\{C,E\}$为例，C 的字母顺序为 3，E 的字母顺序为 5，则其经过散列函数 $h(x,y)=(3\times10+5)\bmod 7$ 余数为 0，所以 2-项目$\{C,E\}$应该放置于第 0 个箱子。在散列函数的选择上应考虑数据库大小，选择合适的函数将交易组合分配于散列表的各对应箱子，以表 3.9 为例，若选择的除数不当，则可能造成过多碰撞（collision），也就是两个项目集在同一箱子中。散列表中的计数值代表该箱子的候选项目集的支持度上限，故若计算结果显示该箱子的支持度未达门槛值时，表示该箱子的所有候选项目集皆非高频项目集，因此即可删除此箱子的所有候选项目集，以提高算法的搜索效率。

表 3.9　散列表形式

箱子位置	0	1	2	3	4	5	6
计数	3	1	2	0	3	1	3
项目集	$\{C,E\}$ $\{C,E\}$ $\{A,D\}$	$\{A,E\}$	$\{B,C\}$ $\{B,C\}$		$\{B,E\}$ $\{B,E\}$ $\{B,E\}$	$\{A,B\}$	$\{A,C\}$ $\{C,D\}$ $\{A,C\}$

相较于 Apriori 算法借由联结上一层级的高频项目集产生新的候选项目集，接着再重新计算这些新候选项目集的支持度，为此须不断搜索整个数据库导致效率不足。DHP 算法则是利用散列树的架构，设计一个散列函数，将数据库中的项目集对应至散列表中，以累计各散列阶层（bucket）所包含项目集的个数；并以所累积的阶层计数粗略估算候选项目集的支持度，以提前删除不可能成为高频项目集的候选项目集。步骤如下：

（1）规定支持度与置信度的门槛值，搜索整个数据库 D 以找出高频 1-项目集 L_1；并且建立 2-项目集的散列表，记为 H_2；定义 $k=1$。

（2）设定 $k=k+1$；利用 L_{k-1} 产生 k-项目集集合 C_k，先利用散列表中各阶层的累积次数来对 C_k 进行初步筛选，再计算筛选后之各 k-项目集支持度以决定高频项目集集合 L_k。

（3）不断地以递归方式重复上一个步骤，直到所有高频项目集集合 L_k 无法再往上一阶层产生 C_{k+1} 为止。

图 3.5 为使用散列表产生候选 2-项目集的范例。使用的数据库 D 如图 3.5(a)所示；首先搜索整个 D，找出候选 1-项目集，也就是 C_1，如图 3.5(b)所示；再依所定的支持度门槛（假设为 0.5）过滤以得第一阶高频项目集 L_1，如图 3.5(c)所示。至此，DHP 算法皆与 Apriori 算法一致，差别在于接下来的二阶项目集搜索。图 3.5(d)定义搜索范围的候选 2-项目集组合，接着利用已给定的散列函数 h 将这些项目集分配至对应的散列箱子中，如表 3.9，以建立二阶散列表 H_2，如图 3.5(e)所示。再比较表 H_2 中的计数与所设定的支持度门槛，假设计数门槛为$\geqslant 2$，以删除计数低于支持度门槛的阶层内的候选 2-项目集，并计算各留下项目集的支持度以求得 2-项目集的候选项目集 L_2，如图 3.5(f)。

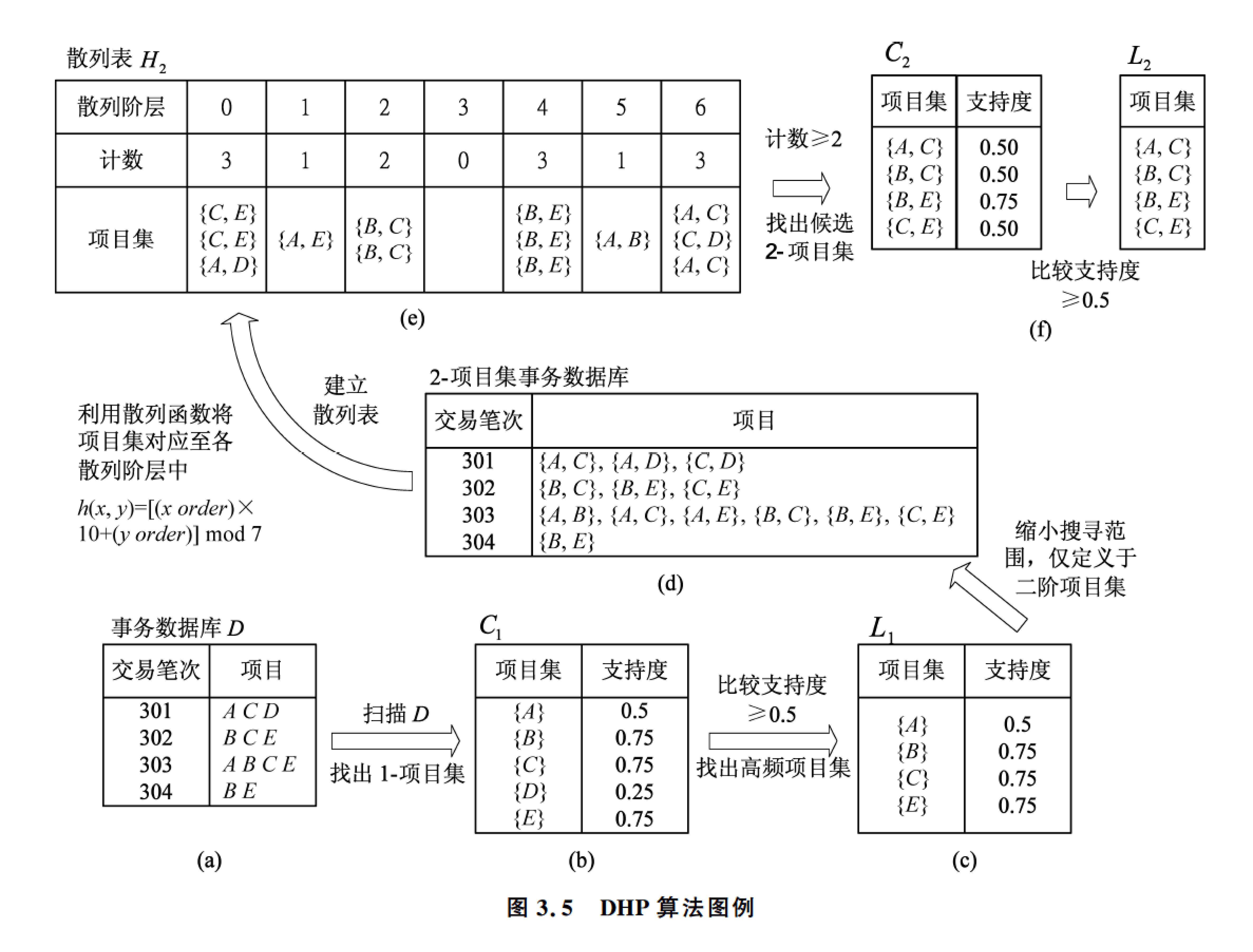


图 3.5　DHP 算法图例

DHP 算法利用散列表的构建来免除大量不必要的低阶(特别是第二阶)候选项目集筛选,其缺点在于一开始必须花费一些时间来建立散列表,且在使用散列阶层所记录的数量来估算候选项目集的支持度时,会使得某些项目集的支持度被高估,而导致初期较高的误判率。然而,只要妥当分析,应可有效地改善后续产生候选项目集的效度。

3.4.4　MSApriori 算法

许多关联规则算法皆假设所有项目或数据变量值出现概率皆为均匀分配,所以都给定固定的支持度门槛以决定高频项目集。然而,实际上,有许多数据项的出现频率并不相同。有时候低频率的项目组合会比高频率的项目组合来得有意义,也会带来较高的效益。因此,刘等(Liu *et al.*,1999)设计一个以 Apriori 为基础的“多重最小支持度关联规则”,称为 MSApriori 算法,提出依不同交易项目,设定多重最小支持度门槛值(multiple minimum supports)的概念,规定每一项目 I_i 的最小支持度 MIS(I_i),若某规则表示为 $I_{i_1},I_{i_2},\cdots,I_{i_k}\Rightarrow I_{j_1},I_{j_2},\cdots,I_{j_l}$,则此规则的支持度只需大于或等于 $\min\{\mathrm{MIS}(I_{i_1}),\mathrm{MIS}(I_{i_2}),\cdots,\mathrm{MIS}(I_{i_k}),\mathrm{MIS}(I_{j_1}),\mathrm{MIS}(I_{j_2}),\cdots,\mathrm{MIS}(I_{j_l})\}$,即具显著性,以处理多重支持度的问题。例如,以商品的购买比例及其所带来的相对效益来决定其支持度门槛值。

在多重最小支持度关联规则中,关联规则的最小支持度为该规则内所有项目集所对应的最小支持度的最小值。分析者对于罕为购买但相对效益高的交易项目(如钻石等)规定了较低的支持度门槛值,对经常购买但相对效益较低的交易项目则规定较高的支持度门槛值

(如牛奶等)。在给予不同门槛值的情况之下,分析者能更合理地找出所要的高频项目集,以产生更客观且符合实际需求的关联规则。

刘等(Liu *et al*.,1999)归纳出关联规则挖掘中多重最小支持度的重要性以及规则特性,称为排序封闭特性(sorted closure property),其概念是由 Apriori 算法的向下封闭的特性延伸而来,即若一项目集满足最小支持度,则该项目集中所有的子项目集也会满足最小支持度,但此特性并不适用于多重最小支持度之关联规则。

假设交易资料库中有四个商品项目,分别记为$\{A\}$、$\{B\}$、$\{C\}$及$\{D\}$。由于交易四种商品所带来的效益不尽相同,所以需给予不同权重。表 3.10 显示分析者对商品项目所规定的**最小支持度门槛值(minimum item support,MIS)**。假设计算出项目集$\{A,B\}$的支持度为 0.08,由于不满足所对应的最小支持度($\min\{0.1,0.2\}=0.1$),因此$\{A,B\}$不属于高频项目集,故不会列入候选项目集中。MSApriori 算法与 Apriori 算法有差异:对于 Apriori 算法而言,若项目集$\{A,B\}$不属于高频项目集,则往上搜索的项目集(如$\{A,B,C\}$或$\{A,B,D\}$)也绝对不会属于高频项目集;但在 MSApriori 算法中,项目集$\{A,B,D\}$的最小支持度为 $\min\{0.1,0.2,0.06\}=0.06$,所以只要项目集$\{A,B,D\}$的支持度大于 0.06,即为高频项目集。换言之,MSApriori 算法不再依循向下封闭的特性来搜索高频项目集,改以排序或权重来搜索候选项目集及建立规则。

表 3.10　各交易项目集的最小支持度门槛值

交易项目	$\{A\}$	$\{B\}$	$\{C\}$	$\{D\}$
MIS	0.1	0.2	0.05	0.06

MSApriori 算法采用多重最小支持度找寻候选项目集并建立显著关联规则,程序如下:

(1) 规定各交易商品项目的 MIS,并将所有交易项目依最小支持度递增排列,而非依循 Apriori 向下封闭的特性。

(2) 先扫描资料库中的所有交易项目,找出符合最小支持度的候选 1-项目集,记为 F_1,并筛选 F_1 以得到高频 1-项目集 L_1。其中,F_1 的每个交易项目都必须在"所有最小支持度的最小值"(即为 minMIS)以上,而 L_1 内的项目都须在"各自的最小项目支持度"以上。

(3) 产生其他候选交易项目集,方法与 Apriori 算法的步骤类似,分为联合(join)与修剪(prune),并以递归的搜索方式依序找出各阶层的候选项目集以及高频项目集。例如,欲产生候选 2-项目集时,必须利用尚未经过最小交易项目支持度测试的项目集集合 F_1 来生成,以避免错失具有效益但出现频率不高的项目集。

图 3.6 为一实际 MSApriori 算法的范例。某事务数据库中有 100 笔商品交易记录,其中包含 4 种商品品项$\{A\}$、$\{B\}$、$\{C\}$及$\{D\}$。经过与专家沟通后所规定的最小支持度门槛值如表 3.10 所示,依照 MSApriori 算法进行高频项目搜索,再依循图 3.6 流程建立关联规则。在第一次扫描数据库后可得到该商品交易 1-项目集组合的支持度如图 3.6(a);在本例中,$\text{minMIS}=\min\{0.1,0.2,0.05,0.06\}=0.05$,所以通过 minMIS 门槛值过滤后的项目集集合 F_1 如图 3.6(b)所示,再经由重新排序后得到图 3.6(c)的项目集 F_1';此时检查 F_1'中各项目集的支持度是否满足其最小支持度后,可得第一阶高频项目集 L_1,如图 3.6(d)所示。接着往上构建候选 2-项目集集合,在此,与 Apriori 算法不同之处在于 MSApriori 算法经由 F_1'来产生 C_2(而非由 L_1 来产生);例如,在此仍保留项目集$\{A\}$来产生 C_2;所构建的 C_2 如

图 3.6(e)所示。接下来重复前三个步骤，借 minMIS 门槛值来删除不会列于候选集合的项目集，产生 F_2 与排序后的 F_2'，分别显示如图 3.6(f)与图 3.6(g)；以各 2-项目集的最小支持度门槛值来删除不满足高频项目集特征的项目组合，可得如图 3.6(h)的高频 2-项目集 L_2。以同样的方式，再往上找出第三阶的候选项目集集合 C_3 及其所对应的高频项目集 L_3，如图 3.6(i)与图 3.6(j)所示，以建立关联规则。

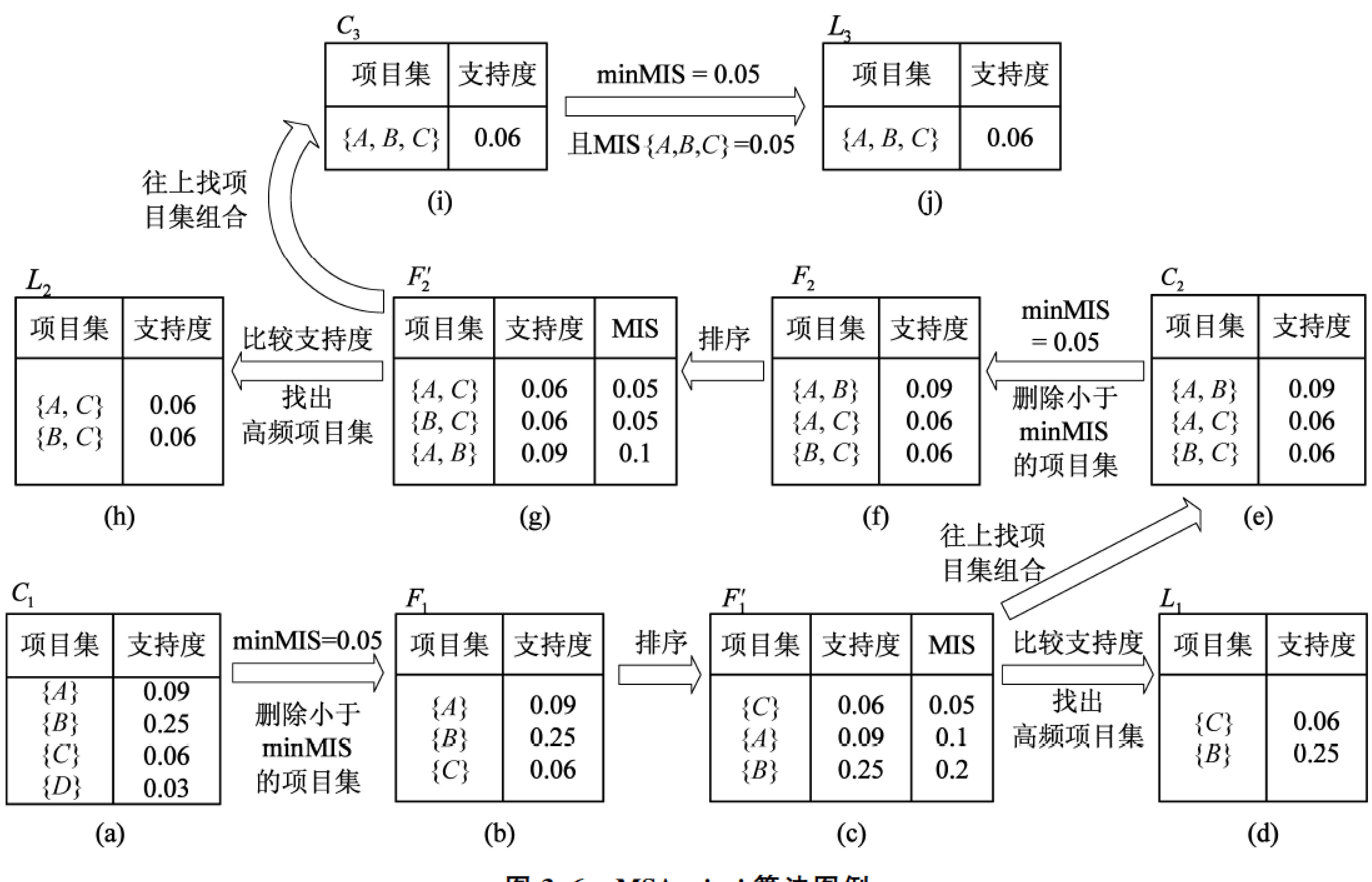


图 3.6　MSApriori 算法图例

MSApriori 算法给予各商品组合不同权重，并依据不同的支持度门槛值来建构关联规则，以避免效益高但发生频率较低的商品组合被删除。如上例，若采用 Apriori 算法，依据向下封闭特性，C_2 必须由 C_1 中支持度在门槛值以上的项目集(即 L_1)所生成，故在候选 2-项目集 C_2 中绝不会包含$\{A\}$项目子集；但是依 MSApriori 算法的概念及符合多重最小支持度的特性，仅有未满足 minMIS 门槛值的项目集(如$\{D\}$)才会被删除而不被用于产生后续更高阶之候选项目集集合。在上例中，项目集$\{A\}$有通过 minMIS 门槛值，所以继续留至候选 2-项目集，如图 3.6(e)所示。

MSApriori 应用相关的机制来避免删除重要但频率较低的项目集，以挖掘频率较低的重要交易规则。然而，MSApriori 的多重最小支持度虽可以找到罕见且重要的规则，但分析者必须对各项商品交易的重要性有一定程度的了解，才能对各项产品项目的最小支持度门槛值做出合适的定义。

3.4.5　FP-Growth 算法

在许多情况下，广度优先搜索算法产生与检查 Apriori 候选项目集合的限制会大幅压缩候选项目集合的大小，并且通常需要产生大量的候选项目集而重复扫描数据库以评估候选项目集的支持度，导致运算效率较低。尽管后续提出许多改善方法，然而在此类架构下所

能提升的效率仍然有限。

频繁模式增长(frequent-pattern growth)算法(简称 FP-增长或 FP-Growth 算法)为目前最有效率的关联规则算法,是将数据库内含有的频繁项目集压缩到一棵频繁模式树(FP-tree)中,并保留项目集之间的重要关联信息。此外,此方法在挖掘时不需产生大量的候选项目集,最多只需扫描数据库两次,因此可大量减少 I/O(input/output)时间,于单一维度及布尔值的领域中,都能以相当有效率的搜索方式建立关联规则(Pei & Han,2000)。

FP-tree 是先储存事务数据库中交易记录项目集所对应的交易记录笔数,并利用相同"前缀"(prefix)共享树中同一路径(path)的原则,将各项目集在数据库出现过的信息紧密压缩储存于 FP-tree 中。由于 FP-tree 主要是用于高频项目集的挖掘,因此树中仅储存各笔交易记录中高频 1-项目集 L_1 所形成的项目集信息,可节省大量的储存空间。FP-tree 的组成为根节点以及每一个交易 1-项目集所代表的叶节点,叶节点中储存了交易项目名称及计数值。

FP-Growth 算法分为两个阶段:第一阶段为建立 FP-tree,第二阶段为挖掘 FP-tree。以表 3.11 某商店的事务数据为例,与专家讨论后规定最小支持度门槛值为 0.6,以下为构建 FP-tree 的三个步骤:

表 3.11　某商店的事务数据库与高频 1-项目集

交易记录	商品交易项目	属于高频项目集并依其支持度大小排序
401	A,B,D,E,F,G	$\{B\},\{A\},\{F\}$
402	B,C,D,F	$\{B\},\{C\},\{F\}$
403	A,B,C,F	$\{B\},\{A\},\{C\},\{F\}$
404	A,B,C,G	$\{B\},\{A\},\{C\}$

(1) 第一次扫描数据库,找出符合最小支持度的第一阶高频项目集,依照支持度大小降序排列,如表 3.11。在扫描数据库后,各项目集的支持度如表 3.12,并与所设定的最小支持度进行比较,以删除不满足门槛值的项目集。过滤后可得出高频 1-项目集集合 L_1,即为 $\{\{A\},\{B\},\{C\},\{F\}\}$,再依其支持度大小排序,得结果为 $\{B\}$、$\{A\}$、$\{C\}$、$\{F\}$,并以该顺序整理原始事务数据,所得结果如表 3.11 最右栏所示。

表 3.12　1-项目集支持度

项目集	$\{A\}$	$\{B\}$	$\{C\}$	$\{D\}$	$\{E\}$	$\{F\}$	$\{G\}$
支持度	0.75	1	0.75	0.5	0.25	0.75	0.25

(2) 建立 FP-tree 的根节点,标识为空节点,然后再次扫描数据库,将属于高频项目集的交易记录依步骤(1)所排列的项目顺序加入 FP-tree 中。进行的方式为先从根节点依序往下搜索是否叶节点已包含欲加入的项目,若已包含则将该叶节点的计数值往上累加;反之,则新增叶节点以储存欲加入的项目。由此可看出 FP-Growth 算法为深度优先搜索算法(top-down),依此种方式继续往下搜索或新增叶节点,直到所有的交易项目名称及出现次数均记录于 FP-tree 中。图 3.7 为根据表 3.11 商店事务数据库所构建的 FP-tree 示意图。

一开始,图 3.7(a)显示产生根节点;而于第二次的扫描中,依照事务数据笔次,第一笔

扫描的数据为{B}→{A}→{F}，因此建立依序产生的叶节点{B}、{A}与{F}及其连结如图 3.7(b)；扫描第二笔项目集{B}→{C}→{F}后，由于树中已有{B}叶节点，故计数值往上加 1，但尚未有{C}叶节点，因此需另建立名称为 C 的叶节点。此外，由于自{C}往下也无{F}叶节点，故需建立新的节点以储存项目{F}。此处需注意的是，不能将图 3.7(c)的{B}→{C}→{F}连结中所新增的{F}叶节点与图 3.7(b)所产生的叶节点{F}视为同一节点，也就是不能建立如图 3.8(b)的节点连结方式，因为若叶节点{F}后续有其他叶节点产生，届时会分不清楚新产生的节点是承接于叶节点{A}或是叶节点{C}的规则。

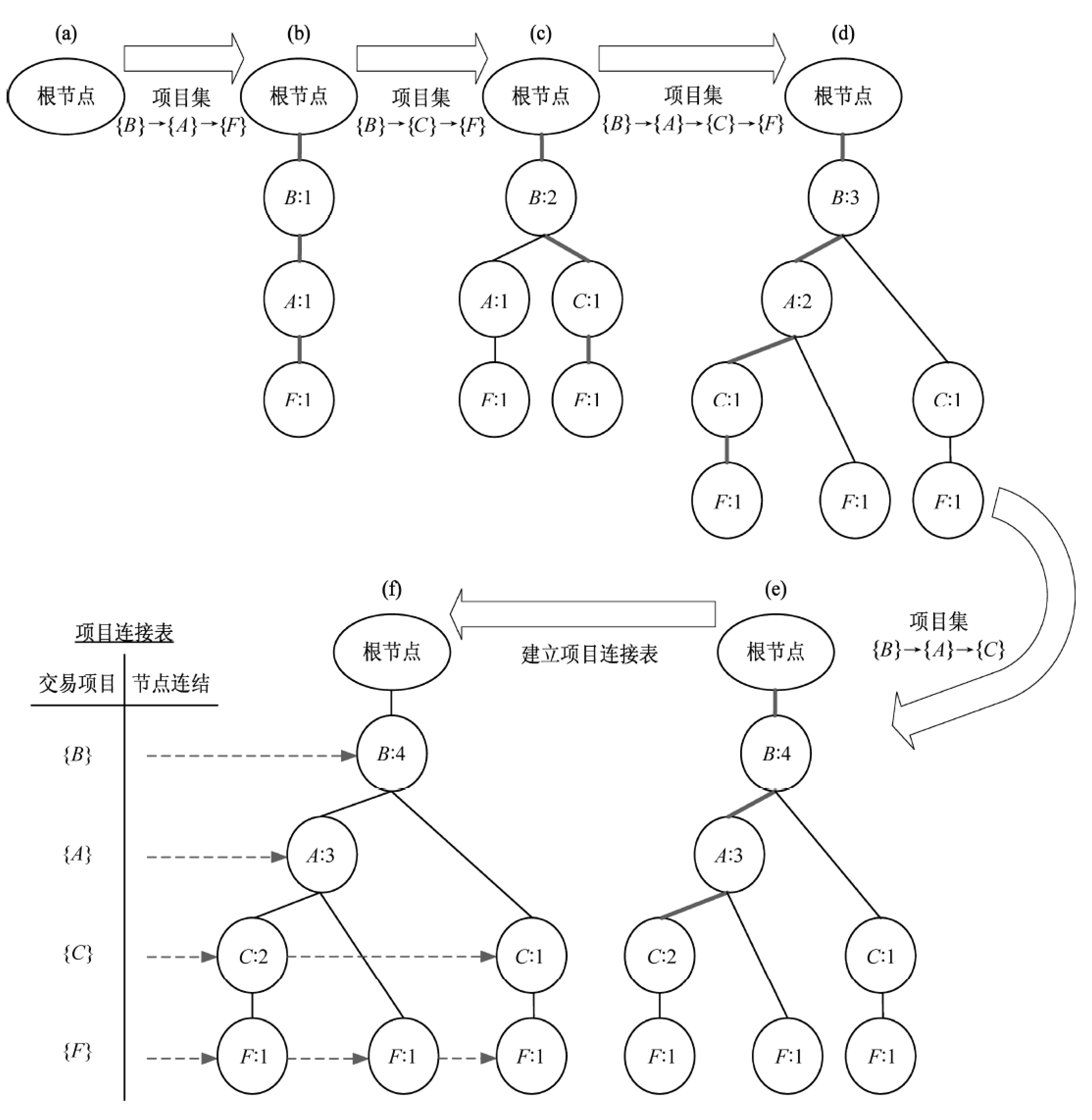


图 3.7　存放已压缩频繁模式信息构建的 FP-tree 示意图

相同地，在扫描完第三笔交易项目集{B}→{A}→{C}→{F}后，在叶节点{B}与{A}的计数值各往上加 1，后续由于叶节点{A}后无叶节点{C}，因此新增一节点以记录项目{C}的事务数据，同样地，叶节点{C}后续也需再增加叶节点{F}，如图 3.7(d)所示；最后，扫描最

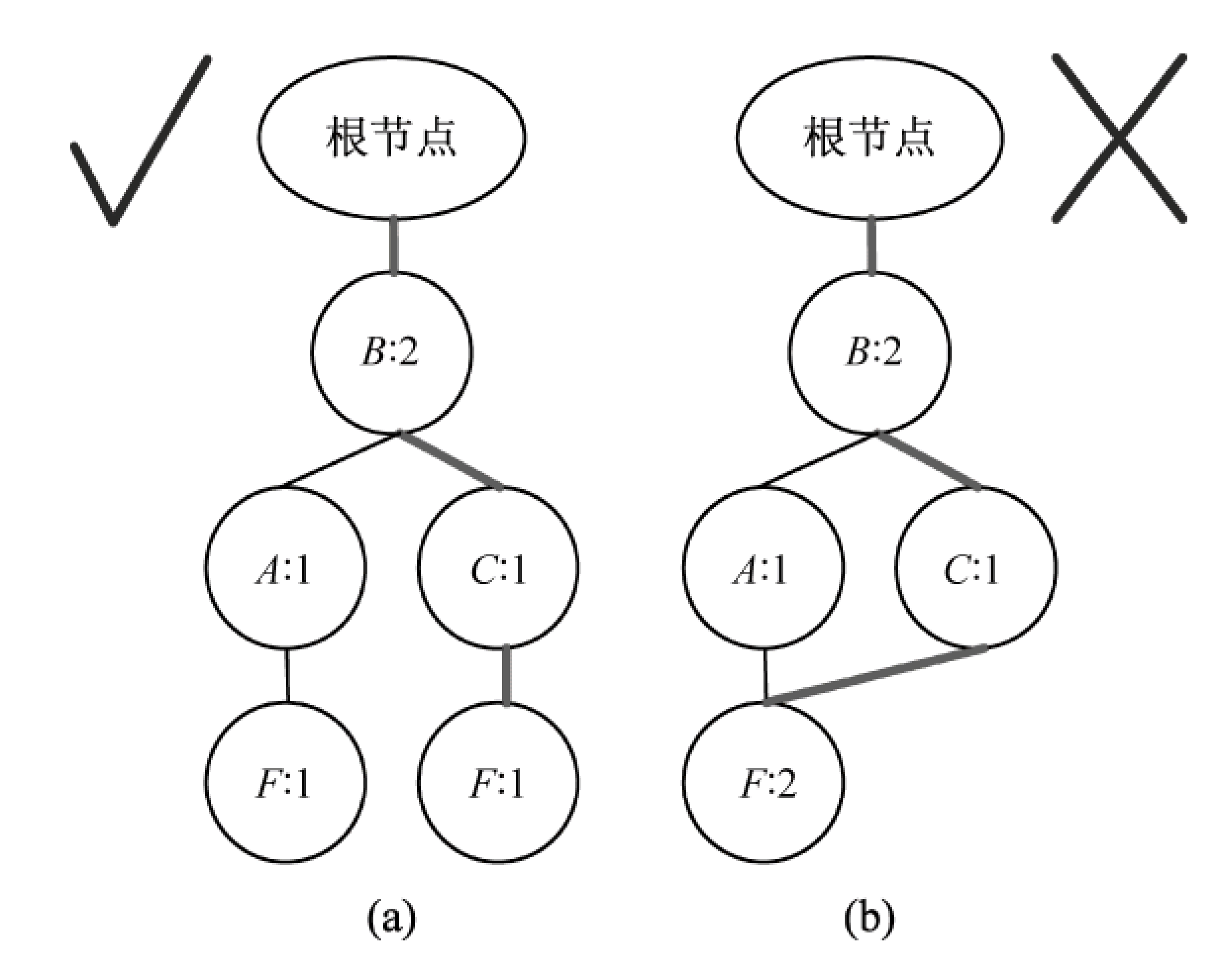


图 3.8 连结节点时所需注意的谬误

后一笔数据建立如图 3.7(e)所示的节点连接形态。由于所有的高频项目集均已扫描过且建立连接关系，至此完成第二次扫描数据库的步骤。

(3) 在此步骤，为了使得 FP-tree 更容易解读，因此建立项目连接表[如图 3.7(f)所示]使每个项目可透过一个节点链来指出该叶节点在树中出现的位置，使树形图更为清晰。项目连接表有两个字段，左边为项目栏，储存高频 1-项目集的项目名称，而右边栏称为横向连结栏，以横向虚线，指出同一项目集于 FP-tree 出现的位置并连接起来，以方便分析人员进一步解读，如图 3.7(f)的虚线所示。

建立完 FP-tree 后，接下来以 FP-Growth 算法针对该树中所隐含的规则进行挖掘，其演算程序分为三阶段：

(1) 由项目连接表中的项目栏由下而上，依叶节点 X 坐落的顺序挖掘，按照每个关联项目连接 FP-tree，以找出 FP-tree 中 X 叶节点的前缀路径，而 X 叶节点的前缀路径所构建的 FP-tree 即称为 X 的条件频繁模式树(简称 X 的条件 FP-tree)。

(2) 以相同方法递归挖掘 X 的条件 FP-tree，计算模式库中每个项目的支持度，找出非空集合且具有高频项目集特征的项目集合，用模式库中的高频项目与 X 组合成高频项目集，列于候选项目集中。可通过前置模式连接 FP-tree 所产生的频繁模式达成模式增长。

(3) 运用阶段一与阶段二的模式不断地对 FP-tree 递归挖掘，找出包含该叶节点的所有前缀路径，直到所有的叶节点均不存在任何前缀路径。

以图 3.7(f)所构建的 FP-tree 为例，说明利用 FP-Growth 算法找出高频项目集的过程。由于是由下往上依照叶节点顺序挖掘，因此以下先探讨包含{F}的所有高频项目集，再依序分别找出包含{C}、{A}以及{B}的高频项目集：

(1) 找出包含{*F*}的所有高频项目集

① FP-tree 的项目连接表中，所有包含{F}的节点计数的加总为 3，表示项目集{F}的支持度为 0.75，大于支持度门槛值 0.6，为高频项目集。

② 从根节点到叶节点{F}包含项目名称 F 的路径有三条，分别为〈{B:4}{A:3}{C:2}{F:1}〉、〈{B:4}{A:3}{F:1}〉及〈{B:4}{C:1}{F:1}〉。由三条路径中，可发现包含项目集{B,A,C,F}、{B,A,F}及{B,C,F}的交易记录皆只有一笔。由{F}的三条前缀路径〈{B}

$\{A\}\{C\}$〉、〈$\{B\}\{A\}$〉以及〈$\{B\}\{C\}$〉所构建的 FP-tree 称为$\{F\}$的条件 FP-tree。

③ 挖掘$\{F\}$的条件 FP-tree，找出〈$\{B\}\{A\}\{C\}$〉、〈$\{B\}\{A\}$〉以及〈$\{B\}\{C\}$〉的交集项目集为$\{B\}$，且由于项目集$\{B\}$支持度为 1，为一高频项目集，因此可和$\{F\}$组成高频项目集$\{B,F\}$。

(2) 找出包含$\{C\}$但不包含$\{F\}$的所有高频项目集

① FP-tree 的项目连接表中，包含$\{C\}$的节点计数加总为 3，表示项目集$\{C\}$的支持度为 0.75，大于支持度门槛值 0.6，为高频项目集。

② 从根节点到叶节点$\{C\}$所包含项目名称 C 的路径有两条，分别为〈$\{B:4\}\{A:3\}\{C:2\}$〉以及〈$\{B:4\}\{C:1\}$〉，第一条路径可看出有两笔交易记录包含项目集$\{B,A,C\}$，而第二条路径则仅有一笔交易记录包含项目集$\{B,C\}$。由$\{C\}$的两条前缀路径〈$\{B\}\{A\}$〉以及〈$\{B\}$〉所构建的 FP-tree 称为$\{C\}$的条件 FP-tree。

③ 挖掘$\{C\}$的条件 FP-tree，所找出〈$\{B\}\{A\}$〉以及〈$\{B\}$〉的交集项目集仅有$\{B\}$，且由于项目集$\{B\}$的支持度为 1，为高频项目集，因此可和$\{C\}$组成高频项目集$\{B,C\}$。

(3) 找出包含$\{A\}$但不包含$\{C\}$和$\{F\}$的所有高频项目集

① FP-tree 的项目连接表中，包含$\{A\}$的节点计数加总为 3，表示项目集$\{A\}$的支持度为 0.75，大于支持度门槛值 0.6，为高频项目集。

② 从根节点到叶节点$\{A\}$所包含项目名称 A 的路径仅有一条，为〈$\{B:4\}\{A:3\}$〉，此路径表示有三笔交易记录包含项目集$\{B,A\}$。由$\{A\}$的两条前缀路径〈$\{B\}$〉所构建的 FP-tree 称为$\{A\}$的条件 FP-tree。

③ 挖掘$\{A\}$的条件 *FP-tree*，由于项目集$\{B\}$支持度为 1，为高频项目集，因此可和$\{A\}$组成高频项目集$\{B,A\}$。

(4) 找出包含$\{B\}$但不包含$\{A\}$、$\{C\}$和$\{F\}$的所有高频项目集

① FP-tree 的项目连接表中，包含$\{B\}$的节点计数加总为 4，表示项目集$\{B\}$的支持度为 1，大于支持度门槛值 0.6，为高频项目集。

② 由项目名称 B 的横向连接找出高频项目集$\{B\}$，由于不存在任何包含$\{B\}$前缀路径，至此结束。

表 3.13 列出采用 FP-Growth 算法所挖掘出的高频项目集合。

表 3.13 利用 FP-Growth 所挖掘出的高频项目集示例

项目集	前缀路径	挖掘出的高频项目集
$\{F\}$	〈$\{B\}\{A\}\{C\}$〉 〈$\{B\}\{A\}$〉 〈$\{B\}\{C\}$〉	$\{B,F\}$
$\{C\}$	〈$\{B\}\{A\}$〉 〈$\{B\}$〉	$\{B,C\}$
$\{A\}$	〈$\{B\}$〉	$\{B,A\}$
$\{B\}$	$\varnothing$	$\{B\}$

FP-Growth 算法以 FP-tree 来储存挖掘的相关信息，将所发现的长频繁模式问题递归地转换成一些短模式问题。当有增加或删除数据库的交易记录时，除非情况特殊，否则 FP-

Growth 算法仅需再次扫描异动的部分，并随之调整 FP-tree 的整体结构，即可使之符合更新后的交易内容，进而挖掘出显著的关联规则。由于不需重新扫描整个数据库因此可大幅节省运算时间及搜索成本。

虽然 FP-Growth 算法在多频繁模式中是比较有效率的方法，但其挖掘结果对管理者或决策者可能太过详细，并且在挖掘过程中需要非常多额外的时间及空间来构建 FP-tree。因此，如何在高层次的频繁模式增长的分析中，归纳出较为低阶的关联规则是另一个需要继续探讨的议题。

现今关联规则的研究多着重在改善算法效率，鲜少研究如何决定最小支持度、置信度等议题。然而，这些支持度门槛与置信度门槛的定义不仅会影响整体算法的效率，亦关系到所寻找的关联规则是否具有意义。若支持度门槛值定得太低，会使分析结果包含过多噪声；但定太高又会误砍重要的信息（Liu *et al.*，1999）。因此应谨慎地决定相关的参数与门槛值，特别是最小支持度。关联规则若能与其他数据挖掘的工具结合（如模糊理论、人工神经网络、决策树等），将可进一步提高所挖掘的规则的准确性，在实务上发挥更大作用。

3.5　多维度关联规则

一般的关联规则分析仅是在单笔交易记录内寻找项目之间的关系，例如，“购买尿布⇒购买啤酒”，其中，尿布与啤酒两项目皆来自同一笔交易记录。若在关联规则挖掘中加入多维度的概念，例如将上述例子加入时间为另一维度因子时，则可挖掘出“顾客周末均会购买尿布⇒购买啤酒”。根据此关联规则，当下次顾客购买尿布与啤酒时，则可推论顾客在下个星期五晚上开始将有很大的机会同样会购买尿布与啤酒。此规则所叙述的交易关联未必包含于同一笔事务数据中，可能为具有时间先后的两笔不同交易记录的关联规则。因此，加入多维度的概念有助于找到多笔交易记录中，项目与项目之间于其他维度的关联规则。

关联规则通常只用单一属性值来描述交易中所记录的项目，也就是从单维度的事务数据集中寻找项目间的关联性。然而，为支持更复杂的商业决策和优化，用户通常需要同时记录多个项目属性值，并设定多个属性值的限制式。因此，多维度关联规则挖掘的相关算法应运而生，使记录项目具有多个属性值，并借由定义数个多维度限制式，寻求不同交易间的关系型法则，以推广至高维度空间的事务数据库。基于多维度的概念，可将数据一笔笔依照其对应的属性维度，置入多维度事务数据库中，如图 3.9 的二维事务数据库所示，以便运用此多维数据库搜索高频项目集，并建立显著关联规则。

图 3.9 为以两属性维度 x 与 y 将某顾客的所有交易记录分割成二维事务数据库，该顾客共有 20 笔消费交易记录，包含了四种交易项目集$\{a\}$、$\{b\}$、$\{c\}$以及$\{d\}$，经由适当的分割后使每一笔交易记录的间隔均等。将每一交易项目依照其发生位置加以区隔，如以 Δ_{x_0,y_0} 表示在维度 x 相隔 x_0 间隔、维度 y 相隔 y_0 间隔的交易项目集。此种表现方式可以清楚地找出不同间隔间交易记录的关联性，如规则“$\Delta_{0,0}(c)\Rightarrow\Delta_{0,1}(a)$”，也就是“若顾客在某一 x 与 y 之下购买项目 c 之后，则会在维度 y 相隔一个间隔、维度 x 相同间隔上购买项目 a”，由图 3.10 可发现该规则的支持度为 4/20。

多维度关联规则中的项目或属性皆会包含两个或两个以上的维度（如时间、购买商品），且由多维度数据挖掘出的关联规则必须能反映不同维度间的关联性，如式(3.5)所示：

T16 *c*	T17 *d*	T18 *a*	T19 *b*, *c*	T20 *a*
T11 *a*	T12 *b*	T13 *b*, *c*	T14 *b*	T15 *b*
T6 *b*, *c*	T7 *d*	T8 *a*	T9 *b*	T10 *a*
T1 *a*, *b*, *c*	T2 *b*	T3 *b*, *c*	T4 *d*	T5 *b*, *c*

维度二：*y*

维度一：*x*

图 3.9　二维数据交易记录示意图

T16 *c*	T17 *d*	T18 *a*	T19 *b*, *c*	T20 *a*
T11 *a*	T12 *b*	T13 *b*, *c*	T14 *b*	T15 *b*
T6 *b*, *c*	T7 *d*	T8 *a*	T9 *b*	T10 *a*
T1 *a*, *b*, *c*	T2 *b*	T3 *b*, *c*	T4 *d*	T5 *b*, *c*

维度二：*y*

维度一：*x*

图 3.10　规则 $\Delta_{0,0}(c)\Rightarrow\Delta_{0,1}(a)$于二维数据库坐落位置

$$A_1=\alpha_1,\cdots,A_n=\alpha_n\Rightarrow B_1=\beta_1,\cdots,B_m=\beta_m,\quad 1\leqslant i\leqslant n,\quad 1\leqslant j\leqslant m \tag{3.5}$$

其中，A_i 及 B_j 皆表示数据的属性，而 α_i 和 β_j 则分别为属性 A_i 及 B_j 的值。

多维度关联规则可顾及多个维度的影响，在考虑到规则的可信度、支持度等的同时，搜索更理想的关联规则，包含提升运算效率以及根据用户定义的样板（template）以找出多维度的显著关联规则两个议题，分述如下：

（1）**提升运算效率**：相对于单维度的关联规则挖掘只需记录数据原本具有的项目，在庞大多维度数据库矩阵中，每个项目皆多了相对的位置关系，如图 3.9 所示。因此，每当扫描数据库时，均需判断这些交易记录里事件发生的相对地址关系及发生次数，造成多维度关联规则所需耗费的时间远高于单维度数据。善加运用数据的特殊结构以节省数据库扫描次数是提升多维度关联规则算法效率的关键，Lu 等（Lu *et al*.，1998）以 Apriori 算法为基础，提出 E-Apriori 和 EH-Apriori 等算法以挖掘数据中的多维度关联规则，并以股市事务数据验证其产生的多维度关联规则作为管理者的决策参考。

（2）**根据用户定义的样板**：Feng 等（Feng *et al*.，1999）提出将用户规定的样板运用于多维度数据挖掘的概念，以提升整体运算效率；用户必须先定义一个或多个想要的模型，模型里可能包含用户有兴趣的事件或事件发生的区间，然后依据此样板进行数据挖掘。由于这些样板限制了关联规则的格式，因此在后续采矿中，分析者只需找出符合这些格式，并满足支持度与置信度门槛值的高频项目集即可，以节省大量的运算时间。例如，分析者欲挖掘事件的发生是否存在“当事件 E 和事件 F 出现在同一时间区隔时，则事件 G 会于两个时间区隔后发生”的规则，即可针对项目集合$\{\Delta_0(E)\}$、$\{\Delta_0(F)\}$以及$\{\Delta_2(G)\}$进行搜索以找出候选项目集合，建立关联规则。

3.6　多阶层关联规则

庞大的数据库常会有数据稀疏的特性，使数据项集无法满足用户设定的支持度门槛值，或很难从中发现真正有用的关联规则。因此，若能将原始数据通过属性的分解及延伸，使数据库的交易记录可用类别关系阶层来呈现，即能针对阶级类别的数据库进行关联规则提取，

提出更有用的潜在信息(Han & Fu,1995)。

一般数据库所储存的交易记录均为低阶的项目集合,如表 3.14 某一流行用品拍卖场所记录的事务数据,包括顾客所购买的商品的原始项目集(如帆布鞋、外套等)。因此,欲挖掘更多概念层级的阶层关联规则,必须先建立概念层级树,运用树状结构表示各阶层类别中的项目集,再往上一阶层汇整成更广义的项目集合。

表 3.14　流行用品拍卖场的事务数据

交易笔次	项　目　集
501	帆布鞋、牛仔裤、短外套
502	篮球鞋、短外套
503	低筒皮鞋、V 领上衣
504	娃娃鞋、长外套、卡其裤

图 3.2 为根据表 3.14 建立的概念层级树,该数据库定义三个阶层的商品交易项目,作为挖掘多阶层关联规则的架构。在定义商品的类别概念分层架构后(如阶层一的分支属性为鞋类以及服饰;阶层二的分支属性则分别为运动鞋、休闲鞋、皮鞋以及上衣、外套、裤子),可通过不同阶层的分支属性找出显著的关联规则。由于数据的稀疏性,原始数据的低阶项目(阶层三,如慢跑鞋、篮球鞋等)较不易满足于支持度与置信度门槛。于是经由概念化阶层的定义,把相关类别的数据往上汇整,使之成为能够代表原始低阶项目的广义集合。例如,数据库中特定鞋款与服饰之间的关系可能很难被发掘,但在提升概念层级树后,很容易即可发现某些鞋类与服饰之间的关联规则。

斯里坎特和阿格拉沃尔(Srikant & Agrawal,1995)针对如何从概念层级树中找出显著的多阶层关联规则,提出运用事务数据表与概念层级树寻找高频项目集以建立关联规则的方法,将出现于概念层级树但未出现于事务数据表的项目集,新增至原始事务数据表中所对应的交易项目里,以产生新的事务数据表。例如,流行用品拍卖场事务数据中的第一笔交易记录有帆布鞋、牛仔裤以及短外套,从概念层级树中可得知帆布鞋隶属于休闲鞋款,故在新的事务数据表中,即可将此笔交易改为含有“休闲鞋、帆布鞋、鞋类”的三个交易项目。接着再利用 Apriori 算法的概念,对于概念层级树进行挖掘,找出显著的多阶层关联规则。

以表 3.14 的流行用品拍卖场的事务数据为例,假设规定最小支持度为 0.5,若不使用概念性阶层的方法,直接以 Apriori 算法找原始事务数据表的关联规则,则可发现表中的所有鞋类(帆布鞋、篮球鞋、低筒皮鞋以及娃娃鞋)个别出现的频率均为 0.25,皆小于支持度门槛值,所以无法产生任何与鞋类相关的关联规则。然而,若使用多阶层关联法进行挖掘,可发现在第一笔交易记录中存在{帆布鞋},以及第四笔交易记录中存在{娃娃鞋}。通过概念层级树,帆布鞋与娃娃鞋皆隶属于休闲鞋款;同理,卡其裤与牛仔裤均隶属于裤子款式。由此,分析者只要往上一阶层(阶层二)搜索即可发现,{休闲鞋与裤子}的支持度为 0.5,满足所规定的门槛值,因此可建立多阶层关联规则“顾客购买休闲鞋⇒购买裤子”。事实上,多阶层关联规则对商业决策或营销策略会有相当大的帮助,可以协助企业提升决策质量和客户满意度。

在多阶层数据库中,无法直接使用各项目集名称来进行复杂的数据库挖掘。因此,GID

(generalized identifier)利用编码的方式,将概念层级树中所包含的原始数据名称项目重新定义并以数值重新编码,以提取阶层概念的关联规则(Han & Fu,1995)。图 3.2 的概念层级树经编码转换成 GID 的概念层级树,如图 3.11 所示。此概念树共分为三层,阶层一存放流行商品种类(以单码转换,分别为 1 与 2);阶层二是储存各流行商品种类下的商品分类(以单码转换,分别为 1、2 与 3);而阶层三为商品本身(以单码转换,分别为 1 与 2)。例如,帆布鞋的 GID 码为{121}、圆领上衣的 GID 码为{211},此种数值化的编码方式,有助于多阶层关联规则的提取。

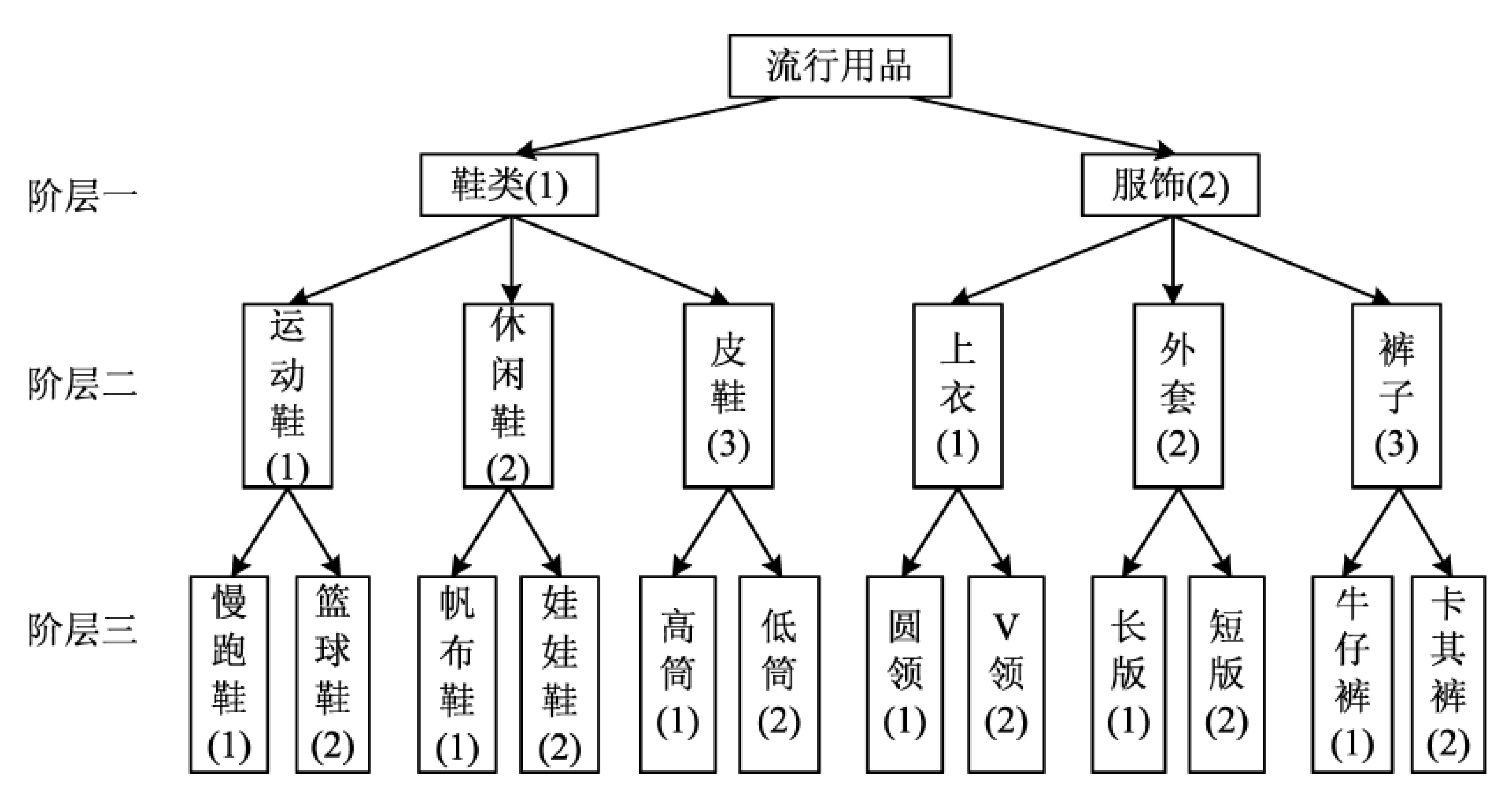


图 3.11 GID 编码后的概念层级树

GID 编码方式主要取决于数据库的交易记录、商品种类以及阶层树的多寡。若于阶层三各商品种类下的商品项目超过十种以上,则可采二位的编码方式,例如,阶层三商品项目慢跑鞋的编码可以{01}表示,则慢跑鞋的 GID 编码即为{1101}。依个别数据库找出适合的编码方式才能使关联规则的建立更有效率。完成 GID 编码后,原流行用品拍卖场的事务数据即可转成以 GID 码储存的交易记录表,如表 3.15 所示。

表 3.15 流行用品拍卖场的事务数据 GID 编码表

交易笔次	项 目 集
501	{121}、{231}、{222}
502	{112}、{222}
503	{132}、{212}
504	{122}、{221}、{232}

多阶层关联规则挖掘的功用与 MSApriori 算法类似,均可避免误删频率低但相对效益高的项目集。两者的最大差别在于前者需定义各种不同阶层的支持度门槛值,但同阶层内各项目的支持度门槛值须一致,以便针对不同层级作关联规则分析。其所产生的关联规则中,高频项目集合内的项目集为低阶项目集的联集,如帆布鞋包含于休闲鞋的项目中;相较之下,后者需定义不同项目集的支持度门槛值,以避免删除相对效益高的项目集合。

3.7　关联规则的应用

数据挖掘要找的是原本不知道但潜藏的有用信息，因此并不是全部符合筛选指针的关联规则皆可拿来应用，必须同时经过领域知识的推论与评估，才能决定哪些规则能够发展成有用的信息。根据实务的解释性，挖掘到的关联规则可区分为两大类，描述如下：

（1）**可依常理推论的规则**：可经由专业领域知识推论，确定为有意义的相关规则。例如"顾客签订维修合约⇒买大型家电用品"、"购买手电筒⇒买电池"、"买桌子⇒买椅子"等。

（2）**巧合造成的无法解释规则**：虽然分析结果显著，但无法由一般常理推导出合理解释的关联规则；此类规则多半为巧合或偶发事件造成，因此无法列为有用的决策参考信息。例如分析结果显示大型五金行的马桶与 A 字梯的销售具高关联性，但两者的关联却无法合理推得，因此无法有效运用该规则。分析者可进一步追踪该现象的成因，或许在某时间点，同时发生两件让民众会购买马桶与梯子的偶发事件，而当初分析所搜集的数据恰好为该段时期。然而这些追踪通常只能了解这些规则的巧合成因，无法应用于日后的销售策略。

关联规则利用分析数据库中各变量与项目集之间的关联性，用于商业实务的应用包括：

（1）**分析顾客行为**：分析客户可能需要哪些服务来提供多样化服务，或是基于消费者购买模式进行相关属性集的数据挖掘，例如，采买商品间的相互关系、年龄与购买行为等。

（2）**进行市场细分与选择目标顾客**：依照关联规则中消费形态将顾客群进行分类以及预测购买行为，以应用于商品货架摆设、库存安排。

（3）**改进卖场陈设与实行目标营销**：将经常一起购买的东西摆在邻近位置，可方便顾客购买；或是将其摆放于购物通道的两端，则可增加顾客寻找商品的滞留时间，促进其他物品的销售量。商品摆放会基于不同类型的商店和卖场经营而异。

（4）**组合搭售商品**：通过消费者购买行为分析，可将顾客会同时购买的相关商品搭配成商品组合以提高销售率；如电信公司提供的套装（捆绑销售）服务。

（5）**发掘诈欺行为**：在反关联规则的交易中，可能存在不合法之行为；例如不寻常的多项保险申请，可能是诈欺行为。

（6）**流失客户分析**：可以分析顾客的流失是否导因于某些关键商品的缺乏等。关联规则分析所得的显著规则因包含高度有效情报，可使公司制订良好的销售策略而提升获利，如因应季节气候变化推出不同的产品组合等。

其他相关的研究与应用包括商业分析、工业技术、医学、生产管理、良率提升与错误检测等。例如，Huang 等（Huang *et al*.，2013）针对健检异常结果与门诊就医记录进行数据的关联规则分析。关联规则的分析结果可挖掘出许多数据库项目之间的联系与相互规则，以作为有用的决策依据，其优点是能从庞大且目标未知的数据库中找出显著性规则、计算模式简单易懂，但当商品数量增加，运算会呈几何级数增加，造成时间耗费，且容易剔除或忽略罕见的商品。

3.8　R 语言与关联规则分析

本节说明通过 R 语言进行 Apriori 关联规则分析，并以 Impact Resources 公司在 1987 年针

对美国旧金山湾区一间购物商场顾客进行问卷营销调查中的部分数据(Hastie *et al*.,2009)为例。此组数据共包含 8993 笔观测值以及 14 个属性,各属性尺度与属性值整理如表 3.16。

表 3.16 范例数据集属性说明

编号	属性名称	数据尺度	属性值
1	income	顺序	[0,10) < [10,15) < [15,20) < [20,25) < [25,30) < [30,40) < [40,50) < [50,75) < 75+
2	sex	类别	male,female
3	marital status	类别	Married,cohabitation,divorced,widowed,single
4	age	顺序	14~17 < 18~24 < 25~34 < 35~44 < 45~54 < 55~64 < 65+
5	education	顺序	grade <9 < grades 9~11 < high school graduate < college (1~3 years) < college graduate < graduate study
6	occupation	类别	professional/managerial, sales, laborer, clerical/service, homemaker,student,military,retired,unemployed
7	years in bay area	顺序	≤1 < 1~3 < 4~6 < 7~10 < ≥10
8	dual incomes	类别	not married,yes,no
9	number in household	顺序	1 < 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9+
10	number of children	顺序	0 < 1 < 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9+
11	householder status	类别	own,rent,live with parents/family
12	type of home	类别	house,condominium,apartment,mobile home,other
13	ethnic classification	类别	american indian, asian, black, east indian, hispanic, pacific islander,white,other
14	language in home	类别	english,Spanish,other

Apriori 关联规则算法的构建与可视化主要应用 R 语言中的 **arules**(Hahsler *et al*.,2014)与 **arulesViz**(Hahsler & Chelluboina,2014)两个扩充套件,而该数据集已内建在 **arules** 扩充套件中。首先,通过以下指令加载扩充套件与数据集:

```
library(arules)
library(arulesViz)
data("IncomeESL")
IncomeESL <- IncomeESL[complete.cases(IncomeESL),]
dim(IncomeESL)
```

删除遗漏值数据后共剩下 6876 笔完整数据,再转换成可用以进行关联规则分析的 **transactions** 对象,亦即每个属性值均转化为单一项目(item),接着,设定最小支持度门槛值为 0.1、最小置信度门槛值为 0.6 以产生关联规则,Apriori 算法共产生 2360 条规则,包含的项目数量从 1 到 6 都有,并绘制三个衡量指标的散布图,以总览所产生的关联规则:

```
rules <- apriori(Income,parameter= list(support= 0.1,confidence= 0.6))
```

```
summary(rules)
plot(rules,method="grouped")
```

关联规则的群组矩阵图(group matrix plot)可总览产生的关联规则中包含哪些项目，进而选取用户可能有兴趣的规则进行详细检查，如图 3.12 所示。图形右方纵向列出所有产生规则的结果项目(right-hand-side，RHS)，如{occupation＝student}、{income＝[0,10)}等，上方横向则是列出群组化的规则条件项目，(left-hand-side，LHS)，矩阵交会的地方则是以圆圈大小代表该群组规则的支持度，颜色深浅代表增益。

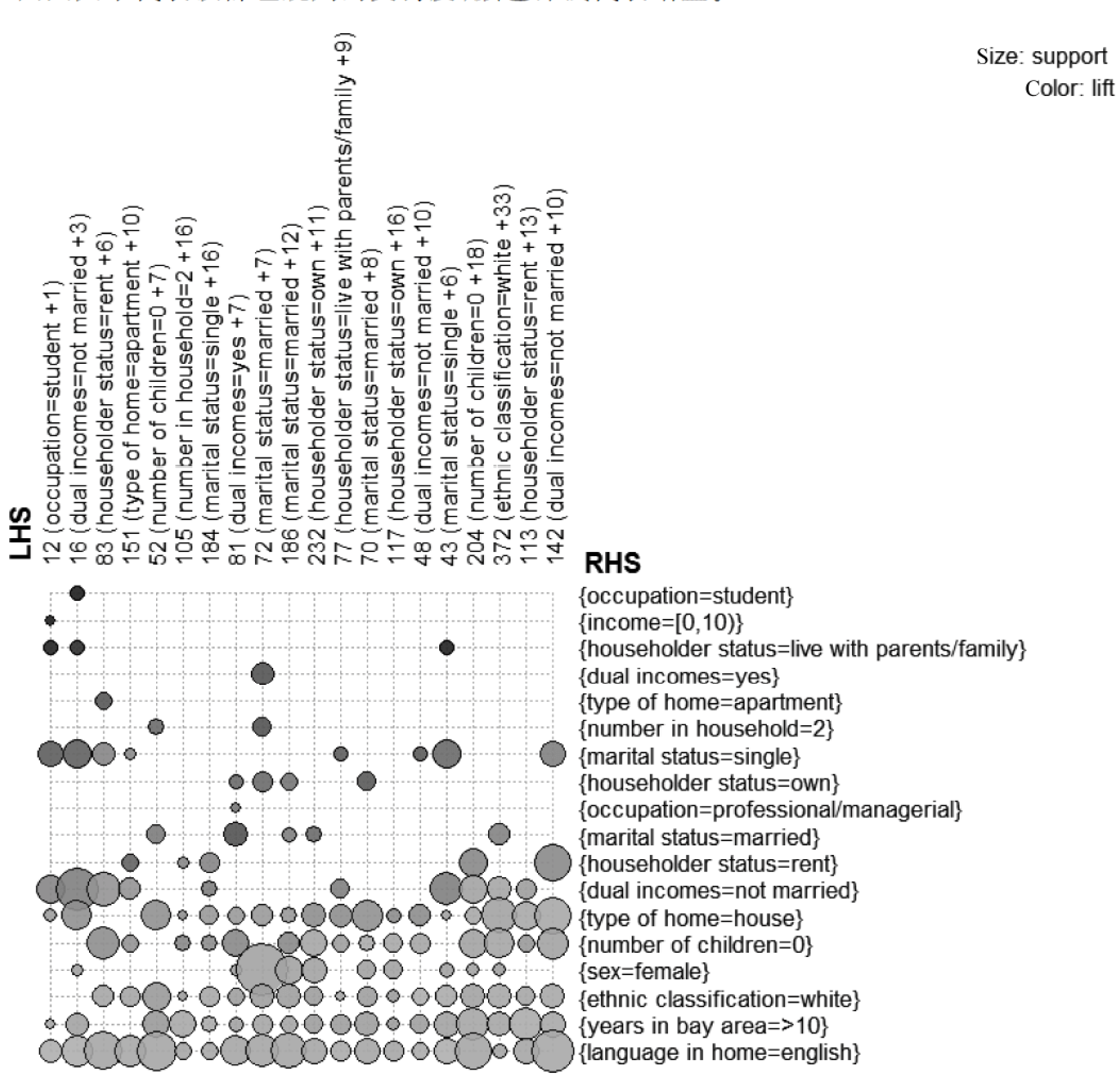


图 3.12　范例数据的关联规则衡量指针散布图

若想了解什么样的人在该旧金山湾区会拥有自己的房子，可以通过筛选 RHS 为{householder status＝own}的显著规则(增益大于 1)，并以支持度排序出前 5 名作进一步检查：

```
rulesOwn <- subset(rules,subset=rhs %in%"householder status=own" & lift>1)
inspect(head(sort(rulesOwn,by="support"),n=5))
```

表3.17整理列出结果为{householder status=own}的显著规则支持度前5名，从中可看出所有规则的条件均有{marital status=married}项目，且单一条件项目的置信度达67.8%，若再加上其他条件项目如{type of home=house}与{language in home=English}，则置信度更可提升至80%以上。

表3.17 限定规则结果下显著规则支持度前5名

编号	条件	结果	支持度	置信度	增益
1	{marital status=married}	{householder status=own}	0.261	0.678	1.804
2	{marital status=married} & {language in home=English}	{householder status=own}	0.247	0.696	1.852
3	{marital status=married} & {type of home=house}	{householder status=own}	0.233	0.828	2.203
4	{marital status=married} & {type of home=house} & {language in home=English}	{householder status=own}	0.221	0.843	2.244
5	{marital status=married} & {ethnic classification=white}	{householder status=own}	0.205	0.735	1.957

若筛选规则时未考虑增益，则有可能得到支持度与置信度都很高，但却无法被采用的规则。通过同样的方式从2360条规则中筛选出使用项目数大于1且增益小于等于1的规则，并以支持度排序得到表3.18。虽然这些规则的支持度与置信度都很高，但由于其结果项目本身都是属于高频项目，若加入条件后的置信度无法高于结果本身的出现频率，便不能算是有效规则。例如，{language in home=English}结果项目本身的出现频率为91.3%，在加入{dual incomes=not married}条件项目后置信度下降至90.7%，代表增加此条件项目对推导{language in home=English}结果项目并无帮助。

表3.18 增益低于1且高支持度与置信度的规则

编号	条件	结果	支持度	置信度	增益
1	{dual incomes=not married}	{language in home=English}	0.543	0.907	0.993
2	{years in bay area=>10}	{ethnic classification=white}	0.430	0.665	0.992
3	{ethnic classification=white}	{years in bay area=>10}	0.430	0.642	0.992

当原始数据中有兴趣的项目并没有相对应的规则产生，此时除了降低产生关联规则的门槛值之外，亦可试着将数据重新编码，降低属性的水平数后再进行关联规则分析。在本范例中，假设对高收入族群有兴趣，但产生的规则中较少有与income结果相关的规则(可参阅图3.12)。此时，可以将原始数据中原本分成9个水平的income属性以$40 000为切点重新编码成“高”与“低”两个水平，并再次进行关联规则分析。

```
library(arules)
library(arulesViz)
```

```
data("IncomeESL")
##remove incomplete cases
IncomeESL <- IncomeESL[complete.cases(IncomeESL),]
##preparing the data set
IncomeESL[["income"]] <- factor((as.numeric(IncomeESL[["income"]])> 6)+1,
levels=1:2,labels=c("$ 40- ","$ 40+ "))
##creating transactions
Income <- as(IncomeESL,"transactions")
#generate rules
rules <- apriori(Income,parameter=list(support=0.2,confidence=0.6))
#screen rules by rhs & lift
rulesIncome <- subset(rules,subset=rhs %in%"income=$ 40+ " & lift>1 )
inspect(sort(rulesIncome,by="confidence"))
```

经过重新转换后的数据在设定最小支持度为 0.2 与最小置信度为 0.6 下共产生 513 条关联规则，其中结果为{income＝＄40＋}的显著规则（增益大于 1）共有 6 条，整理如表 3.19 所示。从中可看出，拥有自己房子{householder status＝own}以及结婚人士为高收入族群的机会较高，置信度在 0.6～0.7 之间。

表 3.19　数据转换后结果为{income＝＄40＋}的关联规则

编号	条　　件	结　　果	支持度	置信度	增益
1	{householder status＝own} & {type of home＝house} & {language in home＝English}	{income＝＄40＋}	0.202	0.676	1.791
2	{householder status＝own} & {type of home＝house}	{income＝＄40＋}	0.211	0.667	1.765
3	{householder status ＝ own} & {language in home＝English}	{income＝＄40＋}	0.233	0.656	1.736
4	{householder status＝own}	{income＝＄40＋}	0.244	0.648	1.717
5	{marital status＝married} & {language in home＝English}	{income＝＄40＋}	0.225	0.633	1.677
6	{marital status＝married}	{income＝＄40＋}	0.237	0.615	1.628

3.9　应用实例——电力公司配电事故定位的研究

3.9.1　案例背景

配电事故对于电力系统的安全性、可靠度以及供电质量均有很大影响。当配电事故发生时，电力公司人员必须检查发生原因或利用发电实验找出事故的发生位置，并进一步将之隔离与维修（Chien *et al*.，2002）。然而一连串的测试与实验，势必会对线路造成某种程度的损害，供电系统亦无法在短时间内修复并恢复作业。因此，如何发展一套可以快速找到事故发生地点的方法来缩短供电恢复时间，为电力公司所关心的议题。

以往一旦发生停电事故，电力公司会立刻派遣人员维修，巡修人员在找到事故发生地点

并进行维修之后,会填写“配电事故停电记录表”,以记录事故发生时现场的相关信息。在过去的维修处理过程中,电力公司累积了大量的配电事故历史数据,每笔数据皆记录有23项属性,如表3.20所示。

表3.20 配电事故记录属性

区 号	总 编 号	馈线代号	停电发生时间	停电总时间
停电用户数	停电电量	气候	停电范围	相数
损坏部位	器材规范	装置年月	制造年月	制造厂
单位	额定容量	事故情形	事故原因	隔离事故设备
分析	环境	电压		

本个案研究(彭金堂等,2005)欲推导事故的损坏设备与预测模式,以求快速找到事故发生地点。因此,属性“损坏部位”为本研究模式的目标项。在其他22项属性中,部分属性涵盖的信息无助于本研究的分析。例如,属性“区号”,由于本个案研究为台北市区,故“区号”编码皆为102,因此可将该属性去除。其他尚有无法在事故发生当下立刻获得的信息,例如“总编号”、“馈线代号”、“停电总时间”、“制造厂”、“分析”、“环境”等在找出事故地点前无法获得的属性,亦暂不纳入考虑,仅留下8项属性数据作为分析模式的输入属性。

另外,在每一属性中,为了避免过多变量造成分析上不必要的复杂性或噪声,因此将“停电时间—月”转为“季节”,并将“停电时间—时”转为“时辰”。最后之输入属性与目标属性整理成如表3.21。

表3.21 分析模式的输入与目标属性

输 入 属 性	目标属性
气候、停电范围、相数、电压、事故情形、事故原因、季节、时辰	损坏部位

3.9.2 数据准备

本个案研究所采用的数据为电力公司于1995—1997年间台北市区的配电事故记录表,共有1649笔数据。首先针对“损坏部位”数据属性进行统计与图表分析,以初步检查数据的分布样型,结果如图3.13所示。

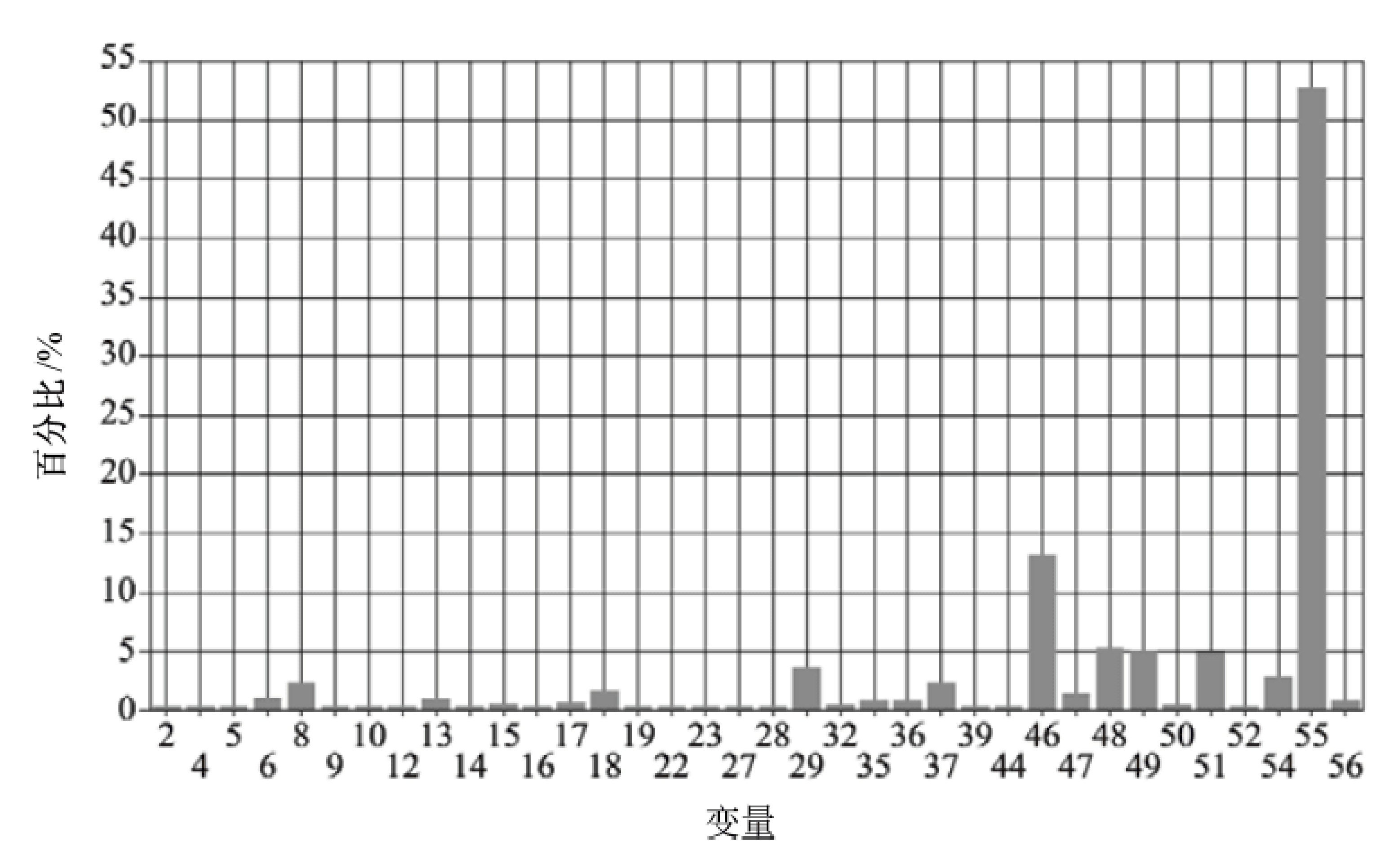


图3.13 “损坏部位”属性中变量分布图

由图 3.13 可知,由于"设备无损坏"(横轴编号 55)占数据笔数 50%以上,相较之下,其他损坏部位项目的相对支持度将非常小。也就是说,在后续分析中许多项目的支持度容易因此显得不够显著,导致忽略这些设备之间损坏的关联规则。整个数据之前置处理过程如图 3.14 所示,在搜集到原始数据后,筛选了 9 项与目标相关之属性(包含输入属性与目标属性),本分析将"损坏部位"属性为"设备无损坏"的数据先行删除再进行分析,以便于察觉其他损坏部位所隐含的信息。然后确认过滤后的数据的完整性。最后得到包含 9 项属性的 780 笔数据,并建立关联规则。

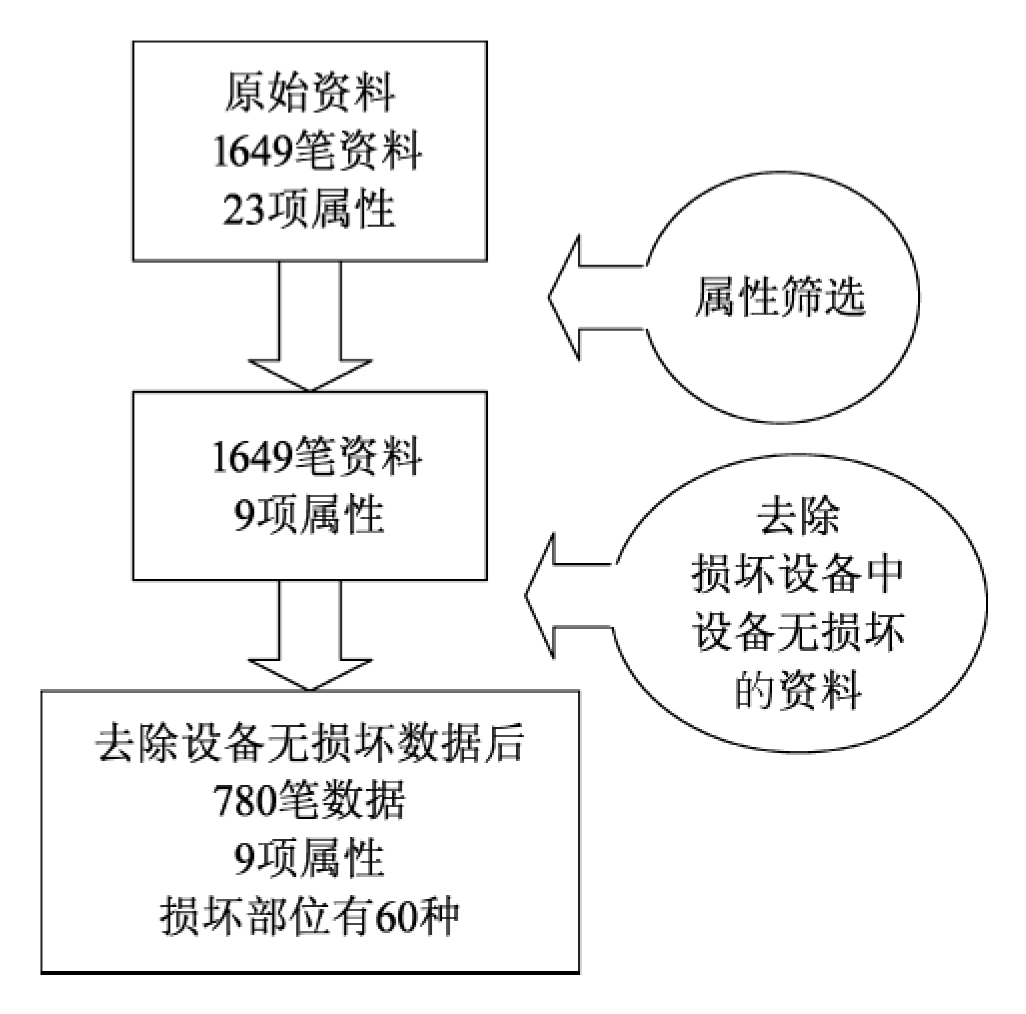


图 3.14　数据前置处理流程

3.9.3　关联规则推导

根据访谈电力公司相关领域的专业人士,本研究设定的最小支持度门槛值为 1.67%,最小置信度门槛值为 50%,而增益门槛值则为 1。参数设定之相关考虑依据如下:

(1) **支持度**:在此数据中,属性"损坏部位"共分为 60 种不同的项目;假设这些项目出现的次数服从项目频率相等的多项分配,则每一损坏项目平均应有 13 笔数据,因此设定与"损坏部位"相关的关联规则支持度应大于 13/780 = 1.67%。由此,将所构建关联规则锁定在频率高于此平均的项目上,故设定支持度门槛值为 1.67%。

(2) **置信度**:依专家经验,掌握线索可推得正确"损坏部位"的概率需大于 50%的规则才具参考价值,因此本研究以 50%为建立关联规则的置信度门槛值。

依据上述参数设定,一共可搜索出 416 条显著的关联规则。由于产生的规则相当多,为了避免噪声太多导致应用上的不便,可依下列步骤(如图 3.15 所示)筛选并删除建立的关联规则,以凸显各目标变量的重要规则。首先,先将所有规则依损坏部位分类,再依置信度递减排序;为避免出现太多无用的规则,根据置信度,选取前 20%的规则。另外,由于许多筛选出的项目集太过冗长,所以最后仅选取输入变量组合长度小于或等于 3 的项目集来建立关联规则。过滤后的关联规则结果如表 3.22 所示,"损坏部位"为"高压电缆"的显著规则有 39 条,为"用户设备"的显著规则有 4 条,而为"高压电缆直线接头"与"熔丝链开关"者则各有 1 条。以"损坏部位"为"高压电缆"的规则 1 为例,当巡修人员发现"事故情形"为"挖断"时,可以推论"损坏部位"为"高压电缆"的概率很高,进而快速采取必要的应对措施。

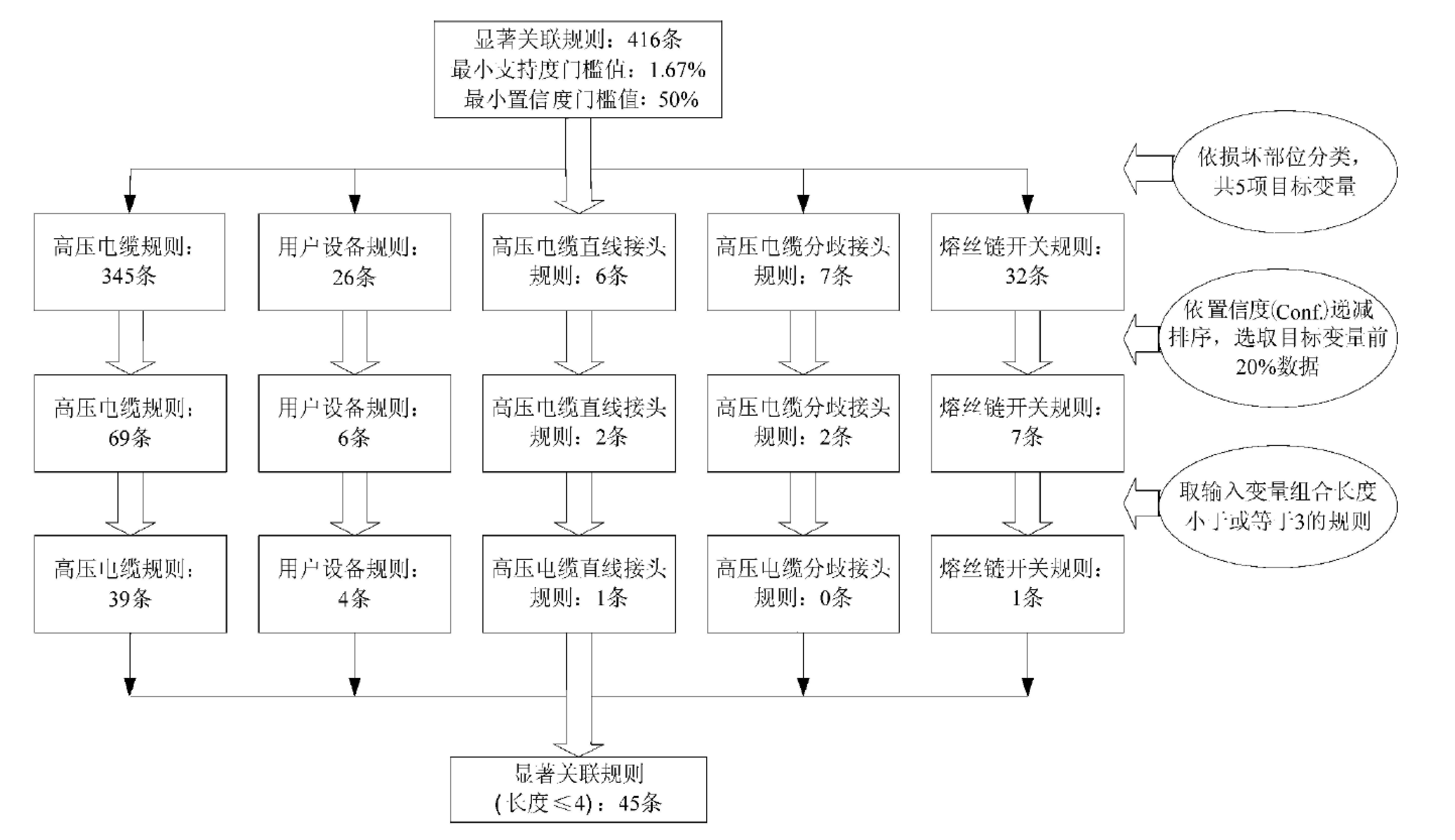


图 3.15 筛选显著关联规则的程序

表 3.22　事故定位关联规则的三项指针

输入变量(前提项目集)	目标变量(损坏部位)	支持度	置信度/%	增　益
1. 事故情形[挖断]	==>[高压电缆]	12.05	100	3.61
2. 事故情形[挖断] 且 事故原因[施工机器碰触]		11.66	100	3.61
3. 事故情形[挖断] 且 季节[5、6、7 月]		3.71	100	3.61
4. 事故情形[挖断] 且 季节[9、10、11 月]		4.23	100	3.61
5. 相数[3ø] 且 事故情形[挖断]		8.33	100	3.61
6. 时辰[9～17 点] 且 事故情形[挖断]		8.20	100	3.61
7. 气候[晴] 且 事故情形[挖断]		8.84	100	3.61
8. 停电范围[地下高压分歧] 且 事故情形[挖断]		6.53	100	3.61
9. 停电范围[地下高压干线] 且 事故情形[挖断]		5	100	3.61
10. 电压[22kV] 且 事故情形[挖断]		6.66	100	3.61
11. 电压[6.6/11.4kV] 且 事故情形[挖断]		5.38	100	3.61
12. 事故情形[挖断] 且 事故原因[施工机器碰触] 且 季节[5、6、7 月]		3.58	100	3.61
13. 事故情形[挖断] 且 事故原因[施工机器碰触] 且 季节[9、10、11 月]		4.10	100	3.61
14. 相数[3ø] 且 事故情形[挖断] 且 事故原因[施工机器碰触]		7.94	100	3.61
15. 相数[3ø] 且 时辰[9～17 点]且 事故情形[挖断]		5.12	100	3.61
16. 相数[3ø] 且 停电范围[地下高压干线] 且 事故情形[挖断]		5	100	3.61
17. 相数[3ø] 且 电压[22kV] 且 事故情形[挖断]		3.97	100	3.61
18. 相数[3ø] 且 电压[6.6/11.4kV] 且 事故情形[挖断]		4.35	100	3.61
19. 时辰[9～17 点] 且 事故情形[挖断] 且 事故原因[施工机器碰触]		7.94	100	3.61
20. 时辰[9～17 点] 且 事故情形[挖断] 且 季节[9、10、11 月]		3.46	100	3.61
21. 时辰[9～17 点] 且 停电范围[地下高压干线] 且 事故原因[施工机器碰触]		3.07	100	3.61
22. 时辰[9～17 点] 且 电压[6.6/11.4kV] 且 事故情形[挖断]		3.07	100	3.61
23. 气候[晴] 且 事故情形[挖断] 且 事故原因[施工机器碰触]		8.71	100	3.61
24. 气候[晴] 且 事故情形[挖断] 且 季节[9、10、11 月]		3.46	100	3.61
25. 气候[晴] 且 相数[3ø] 且 事故情形[挖断]		6.28	100	3.61
26. 气候[晴] 且 时辰[9～17 点] 且 事故情形[挖断]		5.76	100	3.61
27. 气候[晴] 且 停电范围[地下高压分歧] 且 事故情形[挖断]		4.23	100	3.61
28. 气候[晴] 且 停电范围[地下高压干线] 且 事故情形[挖断]		4.10	100	3.61
29. 气候[晴] 且 电压[22kV] 且 事故情形[挖断]		4.61	100	3.61
30. 气候[晴] 且 电压[6.6/11.4kV] 且 事故情形[挖断]		4.23	100	3.61
31. 停电范围[地下高压分歧] 且 事故情形[挖断] 且 事故原因[施工机器碰触]		6.28	100	3.61
32. 停电范围[地下高压分歧] 且 相数[3ø] 且 事故原因[施工机器碰触]		3.46	100	3.61
33. 停电范围[地下高压分歧] 且 相数[3ø] 且 事故情形[挖断]		3.33	100	3.61
34. 停电范围[地下高压分歧] 且 时辰[9～17 点] 且 事故情形[挖断]		4.87	100	3.61
35. 停电范围[地下高压分歧] 且 电压[22kV] 且 事故情形[挖断]		3.97	100	3.61
36. 停电范围[地下高压干线] 且 事故情形[挖断] 且 事故原因[施工机器碰触]		4.87	100	3.61
37. 电压[22kV] 且 事故情形[挖断] 且 事故原因[施工机器碰触]		6.41	100	3.61
38. 电压[22kV] 且 时辰[9～17 点] 且 事故情形[挖断]		5.12	100	3.61
39. 电压[6.6/11.4kV] 且 事故情形[挖断] 且 事故原因[施工机器碰触]		5.25	100	3.61
1. 气候[阴] 且 事故原因[用户设备不良]	==>[用户设备]	2.17	100	17.33
2. 时辰[17 点至次日 1 点] 且 事故原因[用户设备不良]		1.92	100	17.33
3. 相数[3ø] 且 事故原因[用户设备不良]		3.46	100	16.71
4. 相数[3ø] 且 事故原因[用户设备不良] 且 停电范围[高压户]		2.56	100	17.33
1. 电压[22kV] 且 停电范围[地下高压干线] 且 气候[阴]	==>[高压电缆直线接头]	2.30	51.43	4.72
1. 停电范围[架空高压分歧] 且 事故情形[烧损] 且 事故原因[自然劣化]	==>[熔丝链开关]	2.56	71.43	10.13

本个案运用关联规则构建数据挖掘模式，并以电力公司配电事故的历史数据为实证来检验其效度。从架构流程中提取出损坏设备与特殊事故之间的关联模式，提供管理者一具系统化、科学化与量化的参考信息。依照所构建的模式，管理者能根据事故的特定情况来推测出配电事故之样型，以减少事故定位所需的时间。

3.10 结论

关联规则是数据挖掘中最常用于分析顾客交易记录中商品项目关联性的方法之一，亦即从庞大的数据库中，找出数据项集的相关性以建立规则。随着科技进步，数据的快速累积，使数据挖掘在商业与服务业的应用日益受到企业重视。为了获得最大利润、满足客户需求，必须建立良好的顾客关系，以对不同顾客进行服务。以交易数据库为例，每天均有相当大量的消费行为产生，日积月累的数据根本无法通过人脑分析来找出商品销售之间的关联性；因此，若能以适当的演算方法挖掘出不同顾客群的需求，便能发现商机、创造利润。例如，若能知道顾客有同时购买啤酒与尿布的倾向，即可将自制品牌的啤酒与婴儿用品放在一起，以大幅提升获利；同时亦可避免消费者因忘记购买商品而造成的缺憾，提升顾客满意度。

实务上，顾客的消费行为会随着时间而改变，所以需不断地重新挖掘以更新数据库并周期性地执行关联规则运算，以提取出最新的关联规则来洞悉顾客消费形态。因此，发展能大幅减少 I/O 时间的算法对于关联规则挖掘相当重要。除此之外，在产生关联规则的程序中，会产生许多重复或不重要的关联规则，导致所建立的关联规则杂乱无章，因此如何制定合适的支持度、置信度与增益值门槛亦为关联规则分析重要的议题。无论如何，分析者必须对分析数据与所欲达成目标有一定程度的了解，才能选择出恰当的算法，并制定合适的参数指标以构建出有价值的关联规则模式，以提供决策者进行策略决定。

问题与讨论

1. 假设$\{A,B,C\}$为一高频项目集，请列出所有可能由此项目集中搜索出来的关联规则（提示：包括 1-项目集⇒1-项目集、2-项目集⇒1-项目集、1-项目集⇒2-项目集）。

2. 试说明在 Apriori 算法中，支持度与置信度扮演的角色。

3. 下表记录了 24 位患者近两年内的就诊记录，请根据数据回答下列问题。

(1) 请列出所有的 1-项目集，并计算其支持度。

(2) 请列出所有包含项目“糖尿病”的 2-项目集与其所对应的支持度。

(3) 假设支持度门槛为 0.1，请列出支持度高于 0.1 并包含“糖尿病”的所有 2-项目集。

(4) 根据(3)所找出的高频项目集，请计算“糖尿病”对于同项目集中的另一项目的置信度与增益。假设置信度的门槛值为 0.4，在此是否存在任何显著的关联规则？

(5) 请将(2)、(3)、(4)中的项目“糖尿病”分别替换为“贫血”、“高血压”、“忧郁症”、“夜盲症”与“流行感冒”，探讨是否有任何显著的关联规则可被建立。

患者编号	近两年内诊疗记录			患者编号	近两年内诊疗记录		
P01	夜盲症	流行感冒		P13	心肌梗死		
P02	糖尿病	忧郁症	高血压	P14	流行感冒	支气管炎	
P03	忧郁症	流行感冒		P15	糖尿病	流行感冒	
P04	流行感冒	支气管炎		P16	贫血	心脏衰竭	
P05	支气管炎	心脏衰竭		P17	糖尿病	夜盲症	
P06	流行感冒			P18	流行感冒	支气管炎	
P07	糖尿病	心脏衰竭		P19	骨折	夜盲症	
P08	糖尿病	高血压		P20	糖尿病	高血压	
P09	流行感冒			P21	贫血	流行感冒	
P10	贫血	心脏衰竭		P22	糖尿病	高血压	
P11	骨折			P23	贫血	心脏衰竭	
P12	支气管炎			P24	贫血	骨折	

4. 下表为某早餐店所统计的顾客交易记录。请根据数据回答下列问题。

(1) 针对所有顾客的事务数据,请列出所有的 1-项目集,并计算其支持度。

(2) 请找出所有支持度高于 0.2 的 2-项目集。

(3) 请找出所有支持度高于 0.2 的 3-项目集。

(4) 假设支持度门槛为 0.2,置信度门槛为 0.5,请论述"菜包⇒柳橙汁"的规则是否成立?

(5) 承题(4),请论述"烧饼"⇒"油条"、"豆浆"的规则是否成立?

(6) 假设将分析范围锁定为男性顾客,请分别论述(4)和(5)的规则是否成立? 反之,若锁定女性顾客,(4)和(5)规则的成立性又为如何?

(7) 假设将分析范围锁定为 25 岁以上的顾客,请分别论述(4)和(5)的规则是否成立? 反之,若锁定 25 岁以下顾客,(4)和(5)规则的成立性又为如何?

编号	性别	年龄＞25	交易记录
01	男	是	烧饼、菜包、油条、豆浆
02	男	是	菜包、烧饼、油条、豆浆
03	男	是	肉包、菜包、奶茶
04	男	是	烧饼、油条、蛋饼、豆浆
05	男	是	烧饼、油条、豆浆
06	男	否	吐司、蛋饼、奶茶
07	男	否	汉堡、奶茶、可乐
08	男	否	汉堡、油条、豆浆
09	男	否	烧饼、蛋饼、豆浆

续表

编号	性别	年龄>25	交易记录
10	男	否	菜包、煎饺、奶茶
11	女	是	蛋饼、菜包、柳橙汁
12	女	是	油条、豆浆
13	女	是	菜包、肉包、柳橙汁
14	女	是	菜包、柳橙汁
15	女	是	吐司、奶茶
16	女	否	蛋饼、奶茶
17	女	否	菜包、柳橙汁
18	女	否	吐司、松饼、可乐
19	女	否	蛋饼、奶茶
20	女	否	菜包、油条、豆浆

5. 下表为15位受访者的“年龄”、“性别”、“工作产业别”与“薪水”等四项属性的原始数据，分析者欲找出属性之间的关联规则。

(1) 请以“年龄>36”、“年龄≤36”将属性年龄转换成布尔属性，以“薪水>70 000”、“70 000≥薪水>40 000”、“薪水≤40 000”将属性薪水转换成布尔属性，列出其对应的布尔属性值表。

(2) 请根据(1)的布尔属性值表将属性“年龄”与“薪水”类别化，并以0.1为支持度门槛找出所有高频3-项目集。

(3) 假设以0.1为支持度门槛值、以0.5为置信度门槛值，请论述“年龄>36& 性别=女⇒薪水>70 000”的规则是否成立?

(4) 承上题，请论述“年龄>36& 产业别=科技⇒薪水>70 000”的规则是否成立?

(5) 承题(3)，请论述“年龄>36⇒产业别=科技 & 薪水>70 000”的规则是否成立?

编号	年龄	性别	产业别	薪水/元
01	25	男	服务	27 000
02	27	女	学术	38 000
03	28	女	科技	42 000
04	29	女	建筑	45 000
05	29	男	学术	41 000
06	32	女	服务	35 000
07	35	女	科技	53 000
08	37	男	服务	40 000
09	37	男	学术	68 000

续表

编号	年龄	性别	产业别	薪水/元
10	41	女	建筑	42 000
11	42	女	科技	90 000
12	45	男	学术	57 000
13	47	女	科技	100 000
14	51	男	学术	70 000
15	53	男	建筑	81 000

6. 某大型量贩店为了了解顾客消费行为以及产品组合的销售情形，定期搜集各收款机的每笔交易记录，以获得各时段中，各种类型商品的购买次数及单笔数量。下表为抽样五笔交易记录的顾客所购买的商品组合，试回答下列问题：

(1) 利用 Apriori 算法找出所有 2-项目集的可能规则。

(2) 计算所有 2-项目集可能规则的支持度与置信度。

(3) 若支持度门槛值定为 20%，且置信度门槛值定为 20%，试找出被列入候选项目集的规则。

(4) 利用 Apriori 算法找出所有 3-项目集的可能规则、计算其支持度与置信度，并找出被列入候选项目集的规则。

交易记录	商品项目(代码)
601	面包(A)、果酱(B)、花生酱(C)
602	面包(A)、花生酱(C)
603	面包(A)、花生酱(C)、牛奶(D)
604	面包(A)、啤酒(E)
605	牛奶(D)、啤酒(E)

7. 下表为一贩卖 3C 电子产品的连锁店，通过顾客交易记录的搜集、整理以及分析后，所找出的数种可能隐藏信息价值的规则，下表已列出各规则的支持度以及置信度，试计算各规则的增益值并找出候选项目集。

规　　则	支持度	置信度
随身硬盘⇒耳机	60%	75%
耳机⇒随身硬盘	50%	80%
音响喇叭⇒随身硬盘	40%	50%
耳机⇒音响喇叭	30%	40%
音响喇叭⇒耳机	30%	80%
音响喇叭⇒鼠标	10%	10%

8. 给定一高频项目集{A,B,C,D,E,F},则在此项目集之下,最多可能存在多少条关联规则?

9. Partition 算法、DHP 算法与 MSApriori 算法皆为根据 Apriori 算法所衍生出的关联规则搜索方法。请比较此三种算法与 Apriori 算法的差异,并举例说明这些算法的适用性(在什么情况下,这些算法的效果会比 Apriori 算法好)。

10. 试根据以下事务数据,回答下列问题:

(1) 若把每笔交易记录视为一购物篮,试计算商品项目{E},{D,F},{D,E,F}的支持度。

(2) 根据上题结果,计算规则{D,F}→{E}与规则{E}→{D,F}的置信度。

(3) 假设支持度门槛值为 5,试建立该事务数据之 FP-tree。

顾客 ID	交易记录	商 品 项 目
A0341	21 487	D,E
B1254	51 201	A,B,C,D,E,F
A0112	95 481	A,C,E
A0691	61 204	B,D,F
C0387	87 510	A,E
B1254	33 152	C,D,E
A0691	76 541	D,E,F
A0341	22 648	A,F
C0387	15 387	B,E
A0112	01 258	B,C,D

11. 假设一生鲜超市的管理者欲在晚间固定时段将某些隔夜即需丢弃的商品推销售出,并打算以商品合售打折的方式来进行,请问该如何以关联分析来协助策划此方案?再者,若所欲推销的商品占超市内的交易比例不高时,应该以什么算法来进行分析?请详述之。

12. 试举出三个关联规则分析的例子,例如电信公司的促销方案与对应的关联规则应用。

第 4 章

决策树分析

决策树(**decision tree**)具有监督式的特征提取与描述的功能,将输入变量根据目标设定来选择分支变量与分支方式,并以树枝状的层级架构呈现,以提取分类规则。经过修整后的决策树模型可以作为数据探索或预测。决策树可以找出目标变量与各个变量的层级关系。

4.1 决策树的建构

决策树的构建有两个目的:探索与预测,如图 4.1 所示。在决策树探索方面,可以从决策树生长并成形的过程中,由决策树分析结果来解释数据表中隐含的信息,参与决策树生长的数据组仅止于训练数据,待树长成后即可以此探索数据所隐含的信息;在决策树预测方面,可以借由决策树推导的规则来预测未来数据。由于需考虑未来数据进入该模型的分类表现,因此在以训练数据构建决策树后,可应用测试数据来衡量该模式的稳健性与分类表现。通过一连串的验证过程,方可得出最佳的分类规则,作为后续预测之用。例如,简祯富等(2001)针对半导体制程事故诊断,采用决策树分析经过各制程站别间的不同机台路径与测试参数水平之关系,以找出造成产品测试异常的机台设备。

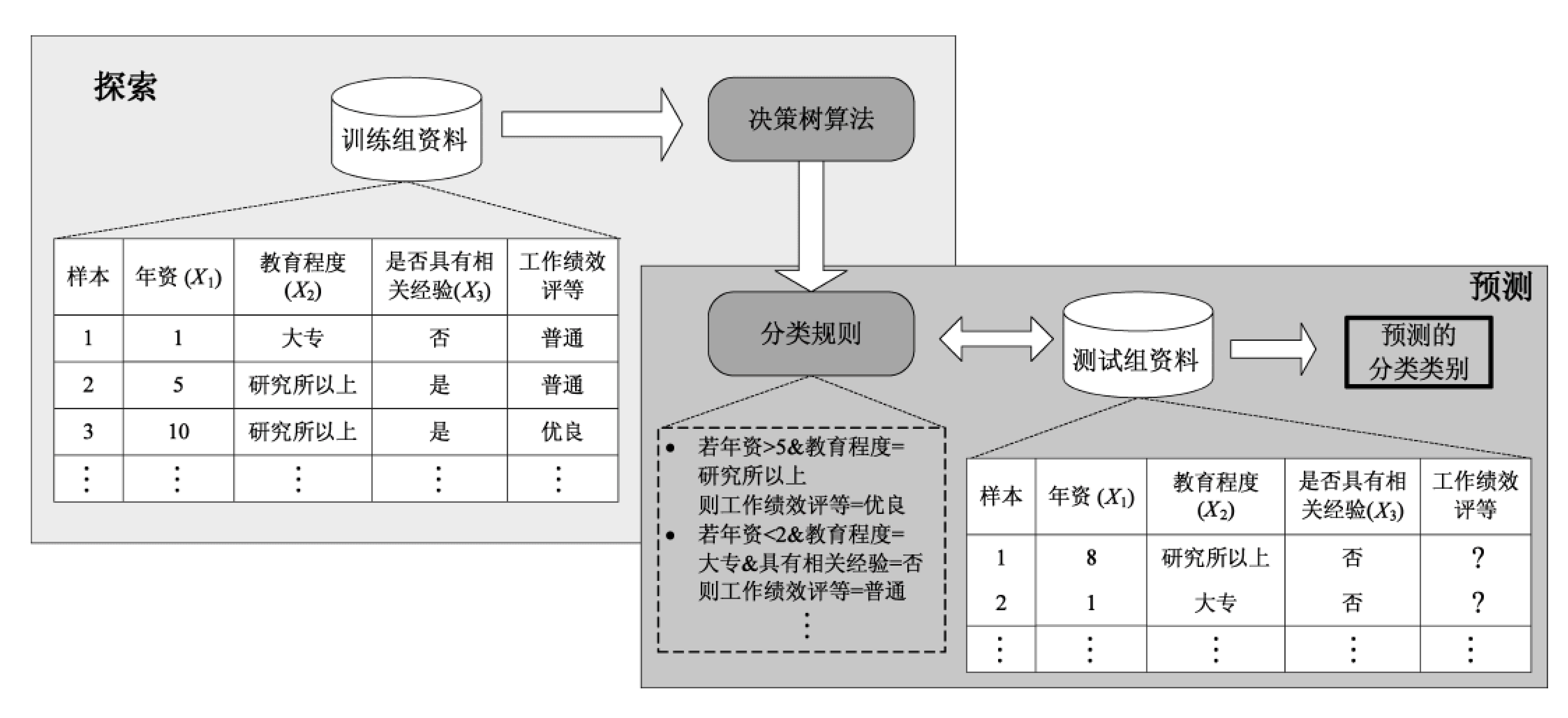


图 4.1 决策树探索与预测

利用决策树进行数据挖掘分析,可将训练组数据放入决策树根部的节点,进行决策树生长的程序,根据问题需求采用适合的算法,包括决定**根节点**(**root node**)以向下分支选择分支变量,并根据分支规则决定根节点的所有数据需进入下一层的哪个内部节点(internal node),

不断重复此分支长树与类别区分，直至所有数据都无法再用显著的分支变量来分类，所有最终层的节点即为**叶节点**(**leaf node**)。当决策树建立完成后，即可将根部到叶节点的每一套独特路径，作为数据分类规则的表达方式。举例来说，以健康为目标建立决策树，若衡量18名人员的血糖(X_1)与血压(X_2)，并以血糖最低标准100与140以及血压最低标准90来做分类，可以发现以下规则：(Ⅰ)若$X_1 \leqslant 100$，则为健康(◎)；(Ⅱ)若$X_1 > 100$且$X_2 \leqslant 140$，则为不健康(▲)；(Ⅲ)若$100 < X_1 \leqslant 140$且$X_2 > 140$，则为健康(◎)；(Ⅳ)若$X_1 > 140$且$X_2 > 140$，则为不健康(▲)。其二维图形与决策树如图4.2所示。

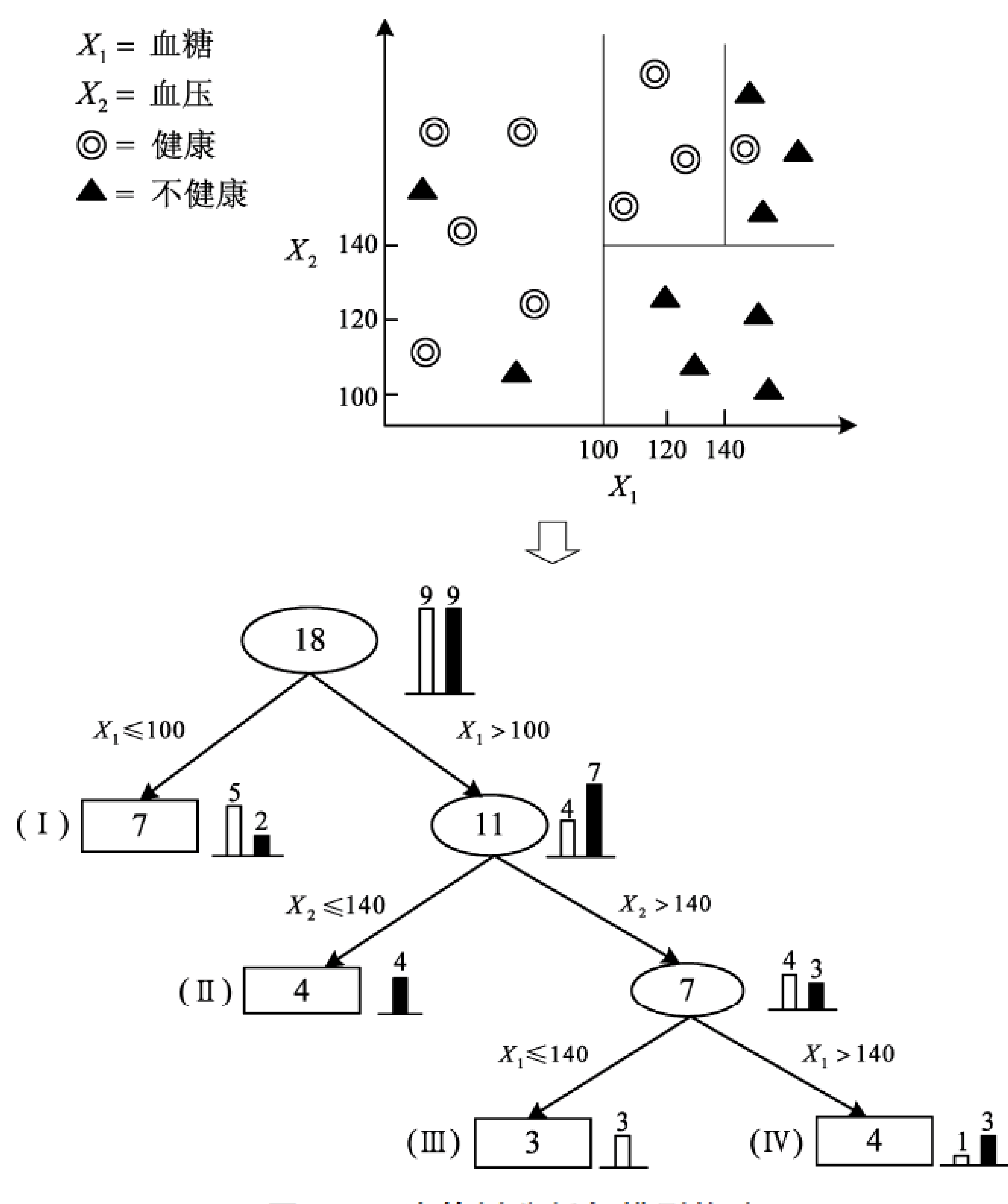


图4.2 决策树分析与模型构建

建立决策树的步骤包括数据准备、决策树生长、决策树修剪及规则提取，如图4.3所示。

4.1.1 数据准备

决策树的分析数据包含两种变量：一为根据问题所决定的**目标变量**；二为根据问题背景与环境所选择的各种属性作为**分支变量**，如图4.4至图4.7所示，分支变量是否容易理解与解释将影响决策树分析结果。

1. **二元属性**：其测试条件可以产生两种结果，如图4.4所示。

2. **名目属性**：名目属性结果的多少可以用不同属性值来表示，例如血型可分为A、B、AB、O四种类别，其分支如图4.5所示。

3. **顺序属性**：可以生成二元或二元以上的分割，其属性可以群组，先决条件是群组必须不违反其属性值顺序特性。例如年龄可分为青年、中年、老年等三种类别，其群组结果如

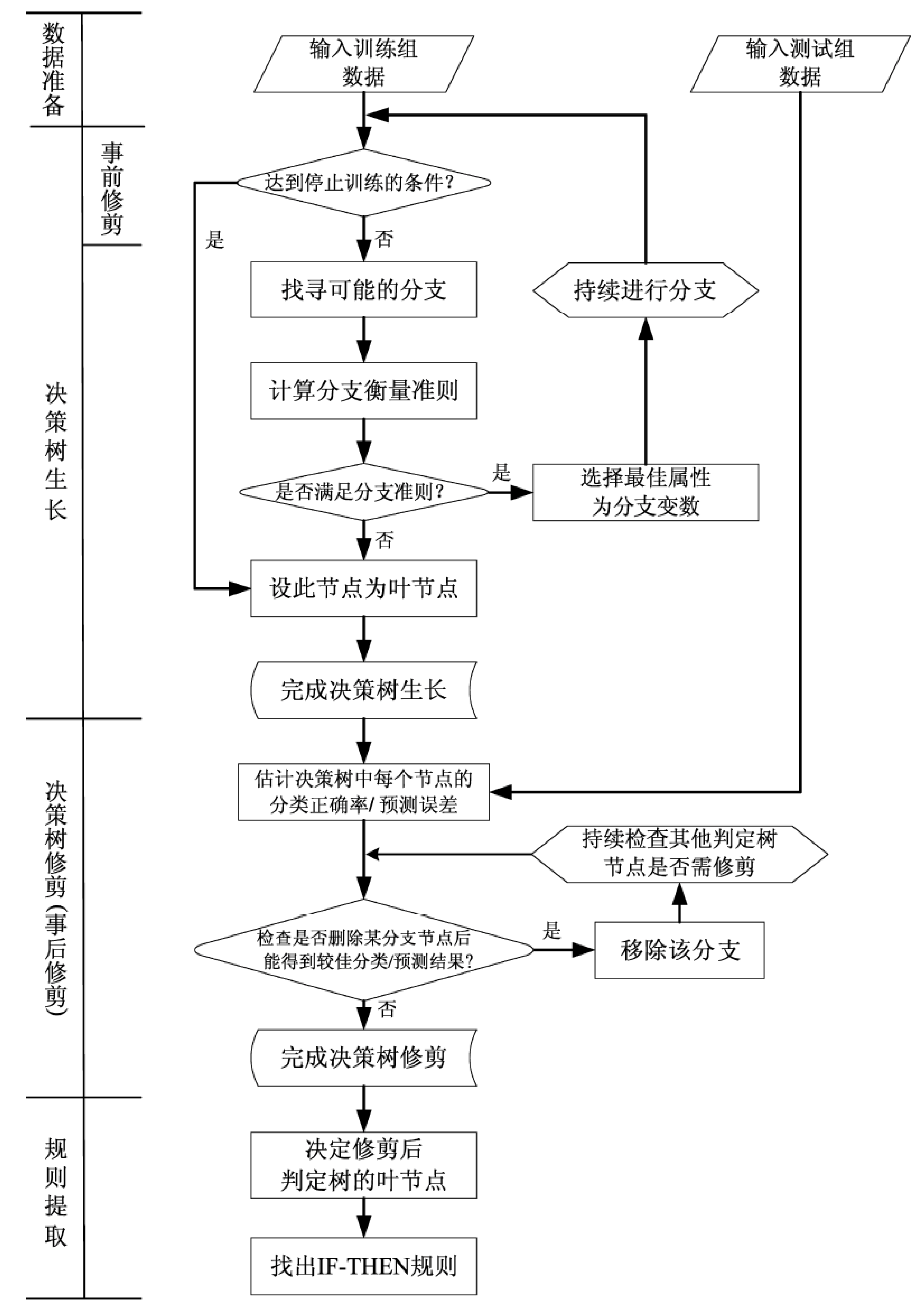


图 4.3　决策树构建的概念步骤

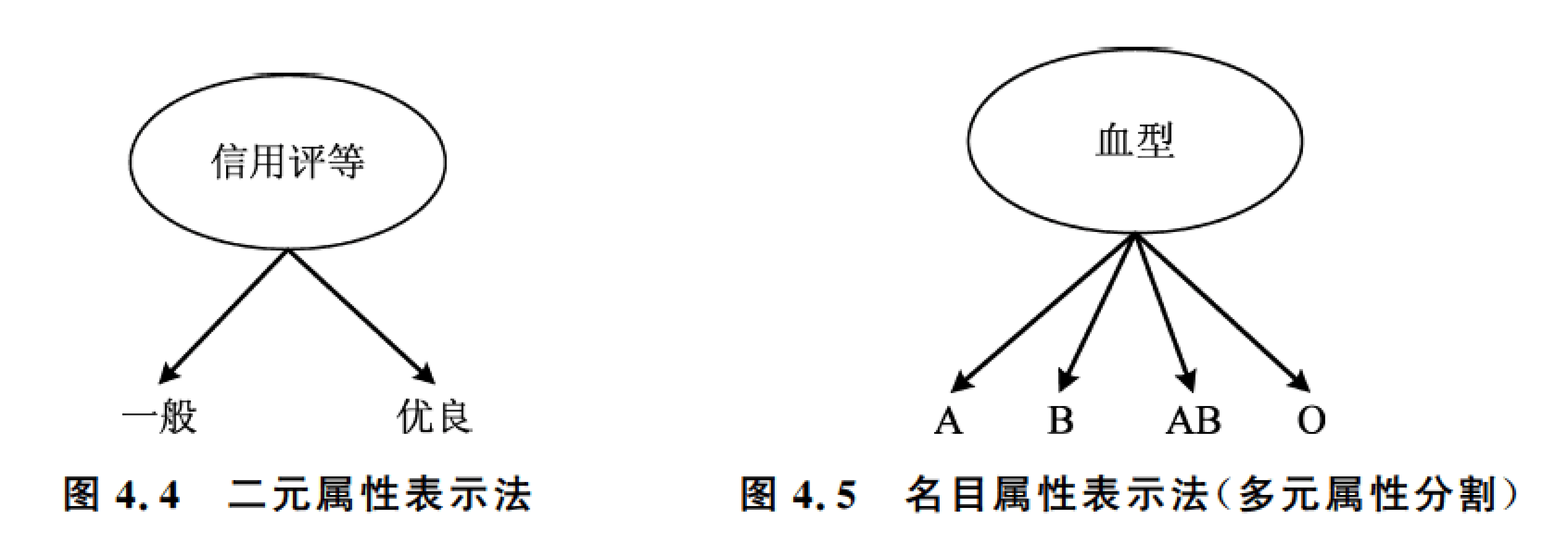


图 4.4　二元属性表示法　　　图 4.5　名目属性表示法(多元属性分割)

图4.6所示;然而,[中年]、[青年,老年]违反了顺序特性,故此顺序不存在。

4. **连续属性**:连续属性的条件可以表示成($X<a$)或($X\geqslant a$)的关系,决策树必须考虑

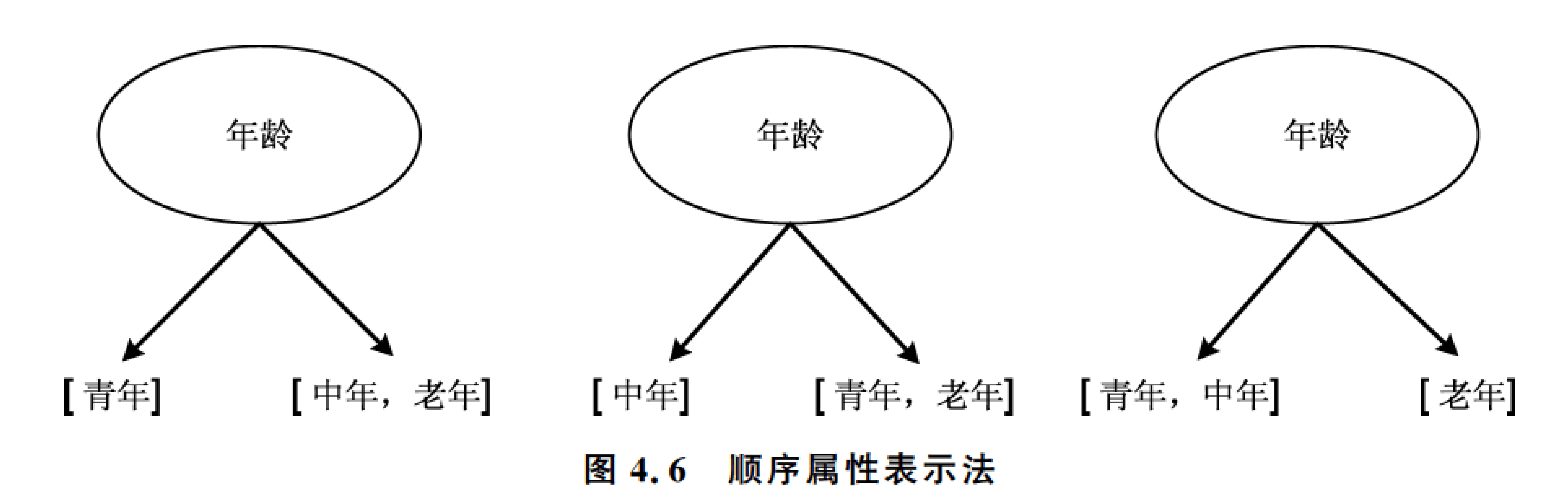


图 4.6 顺序属性表示法

到所有可能的分割点 y，然后再从中选出最好的分割，而在二元以上的分割则须考虑到连续值的范围。在离散化之后，新产生的数值就会分派到指定的区间中，原本相邻的区间也会因此变大，前提要保持顺序性，如图 4.7 所示。

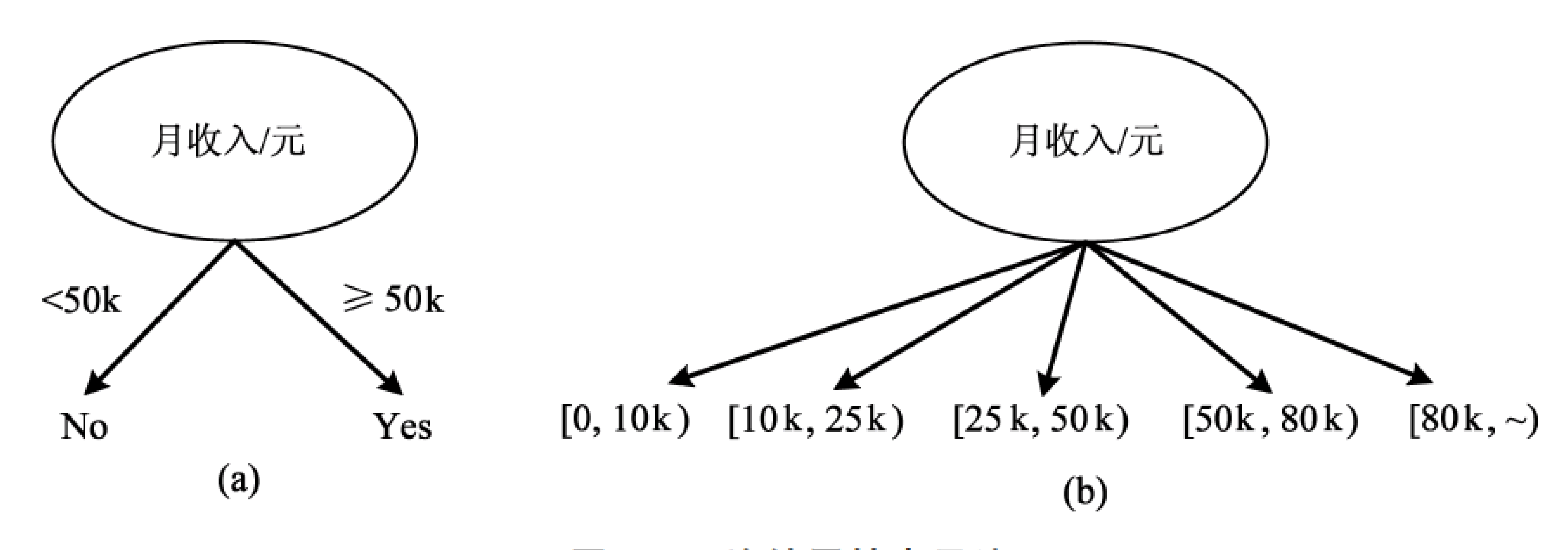


图 4.7 连续属性表示法

取得数据后，再将所搜集的数据分成训练数据集与测试数据集，数据分割详细说明可见第 2 章。前者主要用于决策树模式的构建；后者则用于模式结果的评估。一个好的决策树模式应该能正确分类训练数据集与测试数据集，若一个决策树模式仅在训练数据有很低的错误率，但在测试数据集上却有很高的错误率，则表示该模式过度配适(overfitting)，造成建立的模型无法用于估计其他数据。因此，建立决策树训练模型后，应根据估计测试数据的分类表现，适当地修剪决策树，增加其分类或预测的正确性，并避免过度配适。

4.1.2 决策树的分支准则

决策树的分支准则(splitting criteria)决定树的规模大小，包含树的宽度以及深度。常见的分支准则包括信息增益(information gain)、Gini 系数(Gini index)、卡方统计量(Chi-square statistic)、信息增益比(information gain ratio)等。通过检验分支属性的显著性后，分支准则即能找出具有最佳分支结果的属性。特别的是，在决策树分支过程中，分支属性可以重复出现，亦即各属性有可能使用两次以上，而作为不同层的分支变量。

如表 4.1，假设训练数据集合 D 中有 k 个类别，则 C_j，$j=1,2,\cdots,k$，属性 A 有 l 种不同的数据值。

1. 信息增益

信息衡量(information measurement)是根据不同信息的似然值或概率，以衡量不同条件下的信息量(Quinlan，1983)，如式(4.1)所示。若数据所带来的各种信息的概率皆一致，则获得的信息量亦最大；反之，若各种信息的概率皆不一致，则获得的信息量为最小，而评估函数的价值亦取决于数据所带来的信息状态个数。

表 4.1　决策树分析数据表

属性 A 的数据值 \ 类别	C_1	C_2	…	C_k	总和
A_1	x_{11}	x_{12}	…	x_{1k}	$x_{1\cdot}$
A_2	x_{21}	x_{22}	…	x_{2k}	$x_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	⋮
A_l	x_{l1}	x_{l2}	…	x_{lk}	$x_{l\cdot}$
总和	$x_{\cdot 1}$	$x_{\cdot 2}$	…	$x_{\cdot k}$	N

若每个类别的数据个数定义为 $x_{\cdot j}$，N 为数据集合中所有数据的个数，各类别出现的概率可定义 $p_j = x_{\cdot j}/N$，根据信息论(information theory)可得到各类别的信息为 $-\log_2 p_j$，因此各类别 $C_1, C_2, \cdots, C_k$ 所带来的信息总和 $Info(D)$ 为

$$
\begin{aligned}
Info(D) &= -\frac{x_{\cdot 1}}{N}\log_2\left(\frac{x_{\cdot 1}}{N}\right) - \frac{x_{\cdot 2}}{N}\log_2\left(\frac{x_{\cdot 2}}{N}\right) - \cdots - \frac{x_{\cdot k}}{N}\log_2\left(\frac{x_{\cdot k}}{N}\right) \\
&= -\sum_{j=1}^{k} p_j \cdot \log_2(p_j)
\end{aligned} \tag{4.1}
$$

其中，$Info(D)$ 又称为熵(entropy)，常用以衡量数据离散程度或乱度，可用 $Info(D)$ 作为评估训练数据集合 D 下所有类别的期望信息，当各类别出现的概率相等，则熵值即为 1，表示该分类的信息杂乱度最高。

假设该数据集合 D 要根据属性 A 进行分割，产生共 L 个数据分割集合 D_i，其中，$x_{i\cdot}$ 为各属性值 A_i 下的分割数据总个数，x_{ij} 为属性值 A_i 下且为类别 C_j 的个数，因此，可计算属性 A_i 下的信息 $Info(A_i)$ 如式(4.2)所示：

$$
Info(A_i) = -\frac{x_{i1}}{x_{i\cdot}}\log_2\left(\frac{x_{i1}}{x_{i\cdot}}\right) - \frac{x_{i2}}{x_{i\cdot}}\log_2\left(\frac{x_{i2}}{x_{i\cdot}}\right) - \cdots - \frac{x_{ik}}{x_{i\cdot}}\log_2\left(\frac{x_{ik}}{x_{i\cdot}}\right) \tag{4.2}
$$

属性 A 的信息则根据各属性值下的数据个数多寡决定，如式(4.3)所示：

$$
\begin{aligned}
Info_A(D) &= \frac{x_{1\cdot}}{N}Info(A_1) + \frac{x_{2\cdot}}{N}Info(A_2) + \cdots + \frac{x_{l\cdot}}{N}\mathrm{Info}(A_l) \\
&= \sum_{i=1}^{l}\frac{x_{i\cdot}}{N}Info(A_i)
\end{aligned} \tag{4.3}
$$

至此，**信息增益(information gain)**可以表示为原始数据的总信息量减去分支后的总信息量，如式(4.4)，表示以属性 A 作为分支属性对信息的贡献程度，以此类推可计算出以各个属性作为分支变量所能带来的信息贡献度，比较后可找出具有最佳信息增益的分支属性。

$$
Gain(A) = Info(D) - Info_A(D) \tag{4.4}
$$

[**范例 4.1**]　假设某公司人力资源部门欲了解职员的表现是否受到年资、教育程度、具备相关经验的影响，找出其绩效评等的分类规则，建立人才招募系统的知识法则，以应用于后续的招募程序。首先，搜集该公司员工的相关数据，抽取 10 位现职员工为样本，为方便说明如何计算各项分支准则，将年资属性值分为 3 个区间，分别为 5 年以下、5 年至 10 年、10 年以上，并将教育程度中硕士与博士合并为研究所，转换后的数据如表 4.2。

表 4.2 职员表现的数据(转换后)

职员	年资(A)	教育程度(B)	有无相关经验(C)	员工表现
001	5 年以下	研究所	是	优等
002	10 年以上	研究所	否	普通
003	5 年以下	研究所	是	优等
004	5 年以下	大专	是	普通
005	5 年以下	研究所	否	优等
006	10 年以上	研究所	是	优等
007	5 年至 10 年	大专	否	普通
008	5 年至 10 年	研究所	是	优等
009	5 年至 10 年	大专	否	普通
010	5 年以下	研究所	是	普通

以某公司 10 位职员表现为例,分别根据表 4.2 年资(A)、教育程度(B)、是否有工作经验(C)等三个属性计算出所有种类所带来的信息量总和 $Info(D)$、各属性值所带来的信息量 $Info(A_i)$及信息衡量指针 $Gain$,计算如下:

$$Info(D) = -\sum_{j=1}^{3} p_j \cdot \log_2(p_j) = -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right) = 1.0$$

若选择年资(A)作为分支属性,则其信息增益计算如下:

$$Info(A_{5年以下}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$$

$$Info(A_{5年至10年}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.918$$

$$Info(A_{10年以上}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1.0$$

$$Info_A(D) = \frac{5}{10} \times 0.971 + \frac{3}{10} \times 0.918 + \frac{2}{10} \times 1.0 = 0.961$$

$$Gain(A) = 1.0 - 0.961 = 0.039$$

若选择教育程度(B)作为分支属性,则其信息增益计算如下:

$$Info(B_{大专}) = -\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$$

$$Info(B_{研究所}) = -\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right) = 0.863$$

$$Info_B(D) = \frac{3}{10} \times 0 + \frac{7}{10} \times 0.863 = 0.604$$

$$Gain(B) = 1.0 - 0.604 = 0.396$$

若选择具备相关工作经验(C)作为分支属性,则其信息增益计算如下:

$$Info(C_{是}) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.918$$

$$Info(C_{否}) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.811$$

$$Info_B(D) = \frac{6}{10} \times 0.918 + \frac{4}{10} \times 0.811 = 0.875$$

$$Gain(C) = 1.0 - 0.875 = 0.125$$

因为教育程度的信息增益(0.396)最大，即教育程度作为分支属性能得到较多信息，因此以教育程度作为分支变量。

2. Gini 系数

Gini 系数是衡量数据集合对于所有类别的**不纯度(impurity)**(Breiman *et al.*，1984)，如式(4.5)所示：

$$Gini(D) = 1 - \sum_{j=1}^{k} p_j^2 \tag{4.5}$$

各属性值 A_i 下数据集合的不纯度 $Gini(A_i)$ 如式(4.6)所示：

$$Gini(A_i) = 1 - \left(\frac{x_{i1}}{x_{i\cdot}}\right)^2 - \left(\frac{x_{i2}}{x_{i\cdot}}\right)^2 - \cdots - \left(\frac{x_{ik}}{x_{i\cdot}}\right)^2 = 1 - \sum_{j=1}^{k}\left(\frac{x_{ij}}{x_{i\cdot}}\right)^2 \tag{4.6}$$

属性 A 的总数据不纯度则等于所有属性值分割下的期望平均，如式(4.7)所示：

$$Gini_A(D) = \frac{x_{1.}}{N}Gini(A_1) + \frac{x_{2.}}{N}Gini(A_2) + \cdots + \frac{x_{l.}}{N}Gini(A_l) \tag{4.7}$$

式(4.7)所得之数值即为以属性 A 作为分支属性的不纯度，不纯度越小表示该属性越适合作为分支属性。以此类推可计算出其他属性作为分支变量所能带来的纯度，通过比较即可找出最适合作为分支的属性，如式(4.8)，拥有最大幅度减少不纯度的属性及其分割子集合，则为该决策树分支属性。

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \tag{4.8}$$

以[范例 4.1]为例，分别根据年资(A)、教育程度(B)、是否有工作经验(C)三个属性计算其 Gini 系数如下。

$$Gini(D) = 1 - (0.5)^2 - (0.5)^2 = 0.5$$

$$Gini_{年资}(D)$$
$$= \frac{5}{10}\left[1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right] + \frac{3}{10}\left[1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right] + \frac{2}{10}\left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right] = 0.473$$

$$Gini_{教育程度}(D) = \frac{3}{10}\left[1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2\right] + \frac{7}{10}\left[1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2\right] = 0.286$$

$$Gini_{有无相关经验}(D) = \frac{6}{10}\left[1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right] + \frac{4}{10}\left[1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right] = 0.417$$

$$\Delta Gini(年资) = Gini(D) - Gini_{年资}(D) = 0.5 - 0.473 = 0.027$$

$$\Delta Gini(教育程度) = Gini(D) - Gini_{教育程度}(D) = 0.5 - 0.286 = 0.214$$

$$\Delta Gini(有无相关经验) = Gini(D) - Gini_{有无相关经验}(D) = 0.5 - 0.417 = 0.083$$

由 Gini 系数可知，以教育程度作为分支属性能得到较多信息。

如果将年资属性直接作为分支属性，则表示需找出如“年资 $\leqslant v$”的结果，v 为该连续属性的一个分割点，所有可能的分割点来自于所有的连续值，决定方式为先将所有数值排序，接着选取邻近的两两数据点的中间值作为可能的候选分割点，分割点的评量依据可选用不

同的分支准则，最好的分割点代表可对该连续属性有最好的分支结果。

表 4.3 以年资为例，分别利用信息增益与 Gini 系数准则说明连续属性的分割过程。首先，年资属性值排序后，共有 7 个连续值，每个连续中取其中间作为年资变量的分割点，依据其大于与小于等于的结果将目标变量划分为两组，依序计算年资在不同分割点下的熵，可得当年资以 13.5 作为分割点时，其信息增益为 $Gain(A)=1.0-0.892=0.108$。

若以年资与 Gini 系数为例说明连续属性的分割过程。依据 7 个分割点计算其 Gini 系数，可得当年资以 13.5 作为分割点时，其计算结果为 $\Delta Gini(A)=0.500-0.444=0.056$。

表 4.3　年资以连续属性进行分割

年　资	1		2		4		6		8		12		15
分割点		1.5		3		5		7		10		13.5	
评级		≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	
优秀		1	4	1	4	3	2	3	2	4	1	4	1	
普通		1	4	2	3	2	3	3	2	4	1	5	0	
$Info_{年资}(D)$		1		0.965		0.971		1		1		0.892		
$Gini_{年资}(D)$		0.500		0.476		0.480		0.500		0.500		0.444		

3. 卡方统计量

卡方统计量(χ^2 statistic)以列联表计算两变量间的相依程度，当计算出的样本卡方统计值越大，表示两变量间的相依程度越高，如式(4.9)所示：

$$\chi^2=\sum_{i=1}^{l}\sum_{j=1}^{k}\frac{(x_{ij}-E_{ij})^2}{E_{ij}},\quad E_{ij}=\frac{x_{i\cdot}\cdot x_{\cdot j}}{N} \tag{4.9}$$

其中，E_{ij} 为列联表中第 i 种属性与第 j 种类数据数目的期望值。

以[范例 4.1]为例，分别计算其样本卡方统计值如下。

$$\chi^2(年资)=\frac{(3-2.5)^2}{2.5}+\frac{(2-2.5)^2}{2.5}+\frac{(1-1.5)^2}{1.5}+\frac{(2-1.5)^2}{1.5}+\frac{(1-1)^2}{1}+\frac{(1-1)^2}{1}=0.533$$

$$\chi^2(教育程度)=\frac{(0-1.5)^2}{1.5}+\frac{(3-1.5)^2}{1.5}+\frac{(5-3.5)^2}{3.5}+\frac{(2-3.5)^2}{3.5}=4.286$$

$$\chi^2(有无相关经验)=\frac{(4-3)^2}{3}+\frac{(2-3)^2}{3}+\frac{(1-2)^2}{2}+\frac{(3-2)^2}{2}=1.67$$

表 4.4　列联表数据与期望值

属性：年资

年资＼表现	优秀	普通	总和
(A_1)	3 (2.5)	2 (2.5)	5
(A_2)	1 (1.5)	2 (1.5)	3
(A_3)	1 (1.0)	1 (1.0)	2
总和	5	5	10

属性：教育程度

教育程度 \ 表现	优秀	普通	总和
(B_1)大专以下	0 (1.5)	3 (1.5)	3
(B_2)研究所以上	5 (3.5)	2 (3.5)	7
总和	5	5	10

属性：有无相关经验

有无相关经验 \ 表现	优秀	普通	总和
(A_1) 是	4 (3)	2 (3)	6
(A_2) 否	1 (2)	3 (2)	4
总和	5	5	10

由教育程度的 χ^2 指标(4.286)最大可知，以教育程度作为分支属性能得到有效区分职员绩效评级的结果。

4. 信息增益比

信息增益会选择分支后能降低数据杂乱度的变量，乱度仅考虑到分类错误的比率，并未考虑到候选属性本身所携带的信息，即信息价值的含义。信息增益会倾向找到具有较多属性值的分支变量，假设以顾客编号作为分支变量，因每个顾客编号都仅对单一结果，因此会产生许多分支，且每个分支的乱度皆为 0，因此以顾客编号作为分支变量具有最大信息增益，但解释上却没有任何意义。

信息增益比(information gain ratio)是考虑候选属性本身所携带的信息，再将这些信息转换至决策树，经由计算信息增益与分支属性的信息量的比值来找出最适合的分支属性(Quinlan，1986)，如式(4.10)与式(4.11)所示：

$$GR(A)=\frac{Gain(A)}{Split\ Info(A)} \tag{4.10}$$

$$Split\ Info(A)=-\sum_{i=1}^{l}\frac{x_{i\cdot}}{N}\cdot\log_2\left(\frac{x_{i\cdot}}{N}\right) \tag{4.11}$$

分支变量的属性水平越多，表示使用该变量越容易得到较大的熵，同时亦代表该属性分支特性不显著，因此会倾向选择具有较小熵值的属性为分支变量。而信息增益比的衡量准则倾向于选择具较小熵值的属性，而较不会考虑具有较高的信息增益值 $Gain(A)$ 的属性，特别是当熵值趋近于 0 时；为了避免此种本末倒置的情况发生，故先计算出所有候选属性所带来的平均信息增益值，并仅从具有“高于”平均信息增益值的候选属性中，找出具有最小熵值的属性作为分支变量。

以[范例 4.1]为例，由于三个属性所带来的平均信息增益值为(0.039＋0.396＋0.125)/3＝0.187，而年资属性所贡献的信息增益值低于平均信息增益(0.039＜0.187)，因此可排除年资属性作为分支变量的可能性，故可知以教育程度作为分支属性能得到较多信息。即使纳入年资作为候选属性，从其增益比值的结果也可推得以教育程度作为分支变量

为最佳选择，如下所示：

$$Split\ Info\ (\text{年资}) = -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{3}{10}\log_2\left(\frac{3}{10}\right) - \frac{2}{10}\log_2\left(\frac{2}{10}\right) = 1.485$$

$$Split\ Info\ (\text{教育程度}) = -\frac{3}{10}\log_2\left(\frac{3}{10}\right) - \frac{7}{10}\log_2\left(\frac{7}{10}\right) = 0.881$$

$$Split\ Info\ (\text{有无相关经验}) = -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right) = 0.971$$

$$GR(\text{年资}) = \frac{0.039}{1.485} = 0.026$$

$$GR(\text{教育程度}) = \frac{0.396}{0.881} = 0.449$$

$$GR(\text{有无相关经验}) = \frac{0.125}{0.971} = 0.129$$

由教育程度的信息增益比(0.449)最大可知，以教育程度作为分支属性能得到有效区分职员绩效评级的结果。

5. 方差缩减

当目标变量为连续属性时，则可用方差缩减(variance reduction)作为分支依据。方差是测量数据值与平均值的差异(即该节点内的各笔数据目标值与目标平均值的均方差)，如式(4.12)所示，接下来以某属性进行分支后，检查其分支节点内数据的方差是否比分支前的方差较低。在评估完所有属性进行分支后的方差后，最后再比较候选属性的方差缩减量，并选出具有最大方差缩减量的属性为分支变量。

$$S_t^2 = \frac{\sum_{i=1}^{N_t}(y_{i,t} - \bar{y}_t)^2}{N_t} \tag{4.12}$$

其中，S_t^2 为节点 t 内数据的变异程度，$y_{i,t}$ 为该节点内各样本所对应的相依变量值，$\bar{y}_t$ 为节点 t 中样本所对应的相依变量平均值，N_t 为节点 t 中的样本数。

以[范例 4.1]为例，若将职员表现数据中的绩效评级改为职员的月收入，如表 4.5，可计算其根节点的平均值与方差为($\bar{y}_{t=1}$，$S_{t=1}^2$)＝(47，121.4)。各属性作为分支变量后的方差如表 4.6，发现以年资为分支属性后决策树叶节点的总方差为 16.34，其方差降低的比例最大，因此，年资为第一层决策树分支变量，根据方差降低的准则，最后的决策树见图 4.8。

表 4.5　职员收入的数据

职员	年资(A)	教育程度(B)	有无相关经验(C)	月收入/千元
001	5 年以下	研究所	是	45
002	10 年以上	研究所	否	60
003	5 年以下	研究所	是	42
004	5 年以下	大专	是	39
005	5 年以下	研究所	否	42
006	10 年以上	研究所	是	75

续表

职员	年资(*A*)	教育程度(*B*)	有无相关经验(*C*)	月收入/千元
007	5 年至 10 年	大专	否	40
008	5 年至 10 年	研究所	是	45
009	5 年至 10 年	大专	否	44
010	5 年以下	研究所	是	38

表 4.6　各属性分支后的方差

	年　　资	教 育 程 度	有无相关经验
分支点	[5 年以下 &5 年至 10 年]、[10 年以上]	[大专]、[研究所]	[否]、[是]
方差	0.8×6.36＋0.2×56.25＝16.34	0.3×4.67＋0.7×149.39＝105.97	0.4×62.75＋0.6×160.22＝121.23

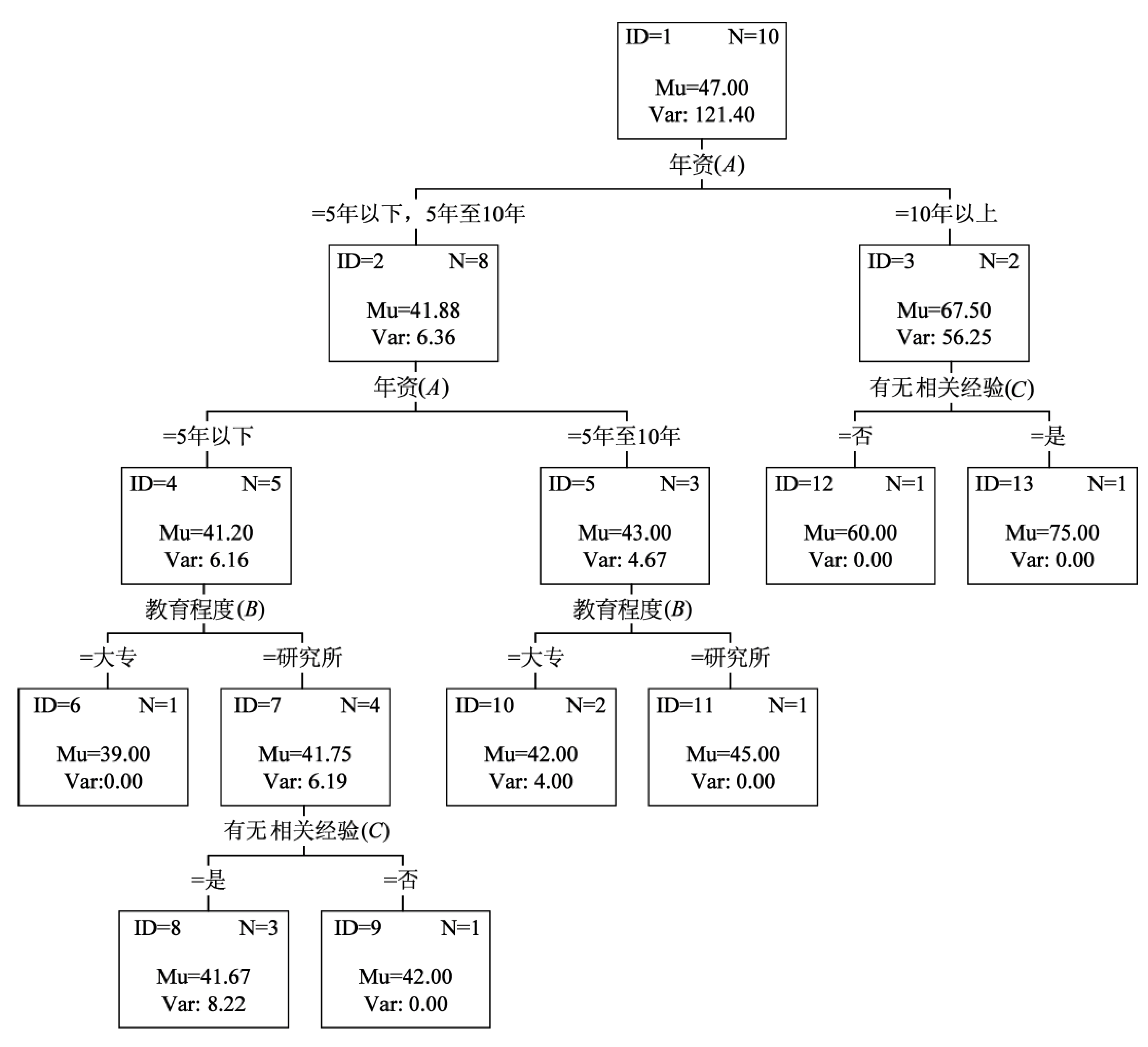


图 4.8　以员工收入为目标变量的决策树

4.1.3 决策树修剪

决策树算法完成树的建立后，各个叶节点即代表不同的种类，部分叶节点可能仅包含少数样本，而没有足够的支持度。决策树生长过程中，树的规模会随着递归的演算方式而扩展，因此一些分支会因为训练数据中隐含的噪声与偏差值而导致过度配适，决策树修剪可以提升未来输入测试数据的预测准确率。修剪决策树的基本原理，即从树的底部开始，检查每个节点和子决策树(sub-trees)，看是否能用一个叶节点替换这棵子决策树，或用其最常使用的分支，可否生成一个分类错误率(classification error rate)更低的子决策树。

决策树修剪方式可分为**事前修剪(pre-pruning)**与**事后修剪(post-pruning)**两种。事前修剪应用于一开始决策树的生长过程中，事先设定停止决策树生长的门槛值，常见的设定门槛如分割的评估值未达此门槛值时，就会停止决策树的生长，例如信息增益值要大于 0.1；或是节点中必须包含足够的样本数目，例如，叶节点中的数据笔数一定要超过 5，则将其标识为叶节点，并停止往下分支。如何决定适当的设定值往往影响最后的结果，太大的设定值常会导致决策树提早收敛，造成解释能力不佳，太小的设定值则会导致决策树过于复杂。事后修剪是在树完全长成后才进行修剪，其引入测试组样本来验证决策树对于新输入数据的分类与预测结果。

事前修剪法的优点在于较具有执行效率，但可能会有过度修剪(over-pruning)的缺点；事后修剪法虽然效率较低，但对于解决决策树的过度配适相当具有正面效益，可避免产生稀少样本数的叶节点，增强决策树对于噪声的忍受程度。

决策树事后修剪方法包括**最小成本复杂修剪(minimal cost-complexity pruning)**(Breiman *et al*.，1984)，同时考虑分类错误率以及决策树的规模大小，先以排列组合的方式列出数种修剪后的决策树，再计算这些树的分类错误率与决策树复杂度(complexity)(即节点个数)，并找出具有最小误差的决策树。

若 $R(t)$代表以节点 t 为起始的决策树的分类错误率，在选择分支时，仅考虑 $R(t)$容易选到较复杂的决策树，因此，对于较复杂的决策树应该给予惩罚，即同时考虑分类错误率与叶节点数目。进行树的修剪时，分类错误率会随着修剪分支的数目呈正比递增，因此，成本复杂性提供降低分类错误率与决策树复杂性之间的权衡方法。如式(4.13)所示，若给定一个复杂系数 α 和未修剪的子决策树节点 t，复杂系数 α 代表的是决策树节点个数的影响，则对某一棵决策树其成本复杂性的定义为决策树节点个数与分类错误率的函数 $R_\alpha(t)$。

$$R_\alpha(t) = R(t) + \alpha \times N_{\text{leaf}} \tag{4.13}$$

其中，$R_\alpha(t)$是该决策树节点 t 造成分类错误率与决策树复杂度的线性组合，N_{leaf}是该决策树中叶节点的数目；$R(t)$为该节点 t 的加权平均分类错误率，也就是该节点的分类错误率与该节点样本数占所有训练样本数比例的乘积。每一个叶节点的分类错误率为节点中无法被正确分类的数据个数占该叶节点全部数据的比例。若有一个节点产生分支，在给定复杂系数 α 下，若分支后的成本复杂度大于分支前的成本复杂度，则进行修剪。

在修剪过程中会产生一连串不同分支的决策树以比较其成本复杂性，对每一个 α，会有一个相对应的子决策树 T_t 将其成本复杂性最小化。当 α 增加时，树的规模会缩小；而当 α 为零时，代表修剪后的决策树与原先决策树的规模相同，具有最小成本复杂性的子决策树将可作为最佳分类与预测的决策树。

成本复杂修剪机制主要有两个步骤：第一步为计算出各子决策树的单位惩罚系数 α，并找出最小 α 以进行树的修剪，第二步则通过验证的方式，输入测试组样本，并从中找出具有最小分类错误率与复杂度的子决策树。

图 4.9 为表 4.2 利用 Gini 系数的完整决策树分支，共有 10 笔样本数、11 个叶节点。可根据类别数据的笔数决定该叶节点的判断类别，以 ID=10 的节点为例，共有 2 笔数据类别为优等，1 笔为普通，所以该节点的预测结果应该判为优等。以图 4.9 中未经修剪的决策树，给定 $\alpha=0.01$ 下，检验节点 8 是否需要修剪，若修剪节点 10 与节点 11，所造成的错分损失成本如下：

$$R(t=8)=\frac{1}{4}\times\frac{4}{10}=0.10$$

图 4.9　职员表现的决策树（未修剪决策树）

然而，若未删除节点 10 与节点 11，在叶节点个数 $N_{leaf}=2$，其产生的加权平均分类错误率如下：

$$R_\alpha(t=8)=\left(\frac{1}{3}\times\frac{3}{10}+0\right)+0.01\times 2=0.12$$

因为 $R_\alpha(t)>R(t)$，所以进行修剪。同样地，给定 $\alpha=0.01$ 下，以节点 5 为例，$R(t=5)=\frac{1}{5}\times\frac{5}{10}=0.10$，$R_\alpha(t=5)=\left(\frac{1}{4}\times\frac{4}{10}+0\right)+0.01\times 2=0.12$，$R_\alpha(t)>R(t)$，所以修剪节点 8 与节

点 9。再以节点 3 为例，$R(t=3)=\frac{3}{7}\times\frac{7}{10}=0.30$，$R_{\alpha}(t=3)=\left(\frac{1}{2}\times\frac{2}{10}+\frac{1}{5}\times\frac{5}{10}\right)+0.01\times 2=0.22$，$R_{\alpha}(t)<R(t)$，所以停止修剪决策树，如图 4.10 所示。

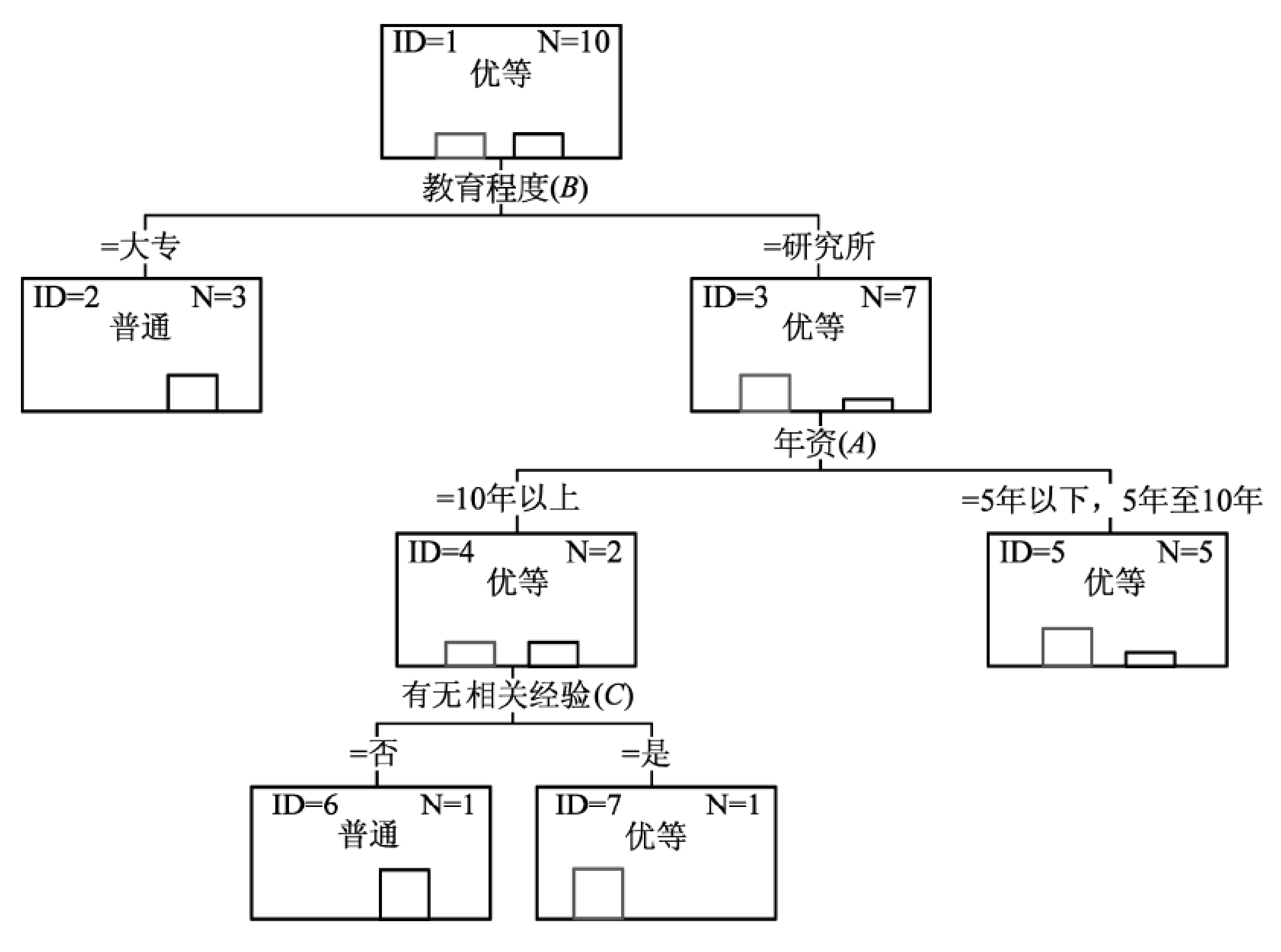


图 4.10　职员表现的决策树(修剪后决策树)

4.1.4　规则提取

完成决策树的生长及修剪后，即可利用决策树提取数据中隐含的信息。**IF-THEN** 规则即为从根节点至叶节点的可能路径(path)。沿着可能路径可串连起作为分支变量的属性，形成一套具因果关系的分类模型，用以分类数据。例如，在图 4.11 笔记本电脑价格的决策树模型示例中，其目标变量为笔记本电脑价格，属性有 CPU 转速以及硬盘容量，通过已建立的决策树可提取出笔记本电脑价格的决策规则如表 4.7 所示。

表 4.7　笔记本电脑价格的决策规则

IF	THEN
若“CPU 速度慢”，且“硬盘容量小”	笔记本电脑的价格是“便宜”
若“CPU 速度慢”，且“硬盘容量大”	笔记本电脑的价格是“中等”
若“CPU 速度中等”，且“笔记本电脑重量为 1kg”	笔记本电脑的价格是“昂贵”
若“CPU 速度中等”，且“笔记本电脑重量为 2kg”	笔记本电脑的价格是“中等”
若“CPU 速度中等”，且“笔记本电脑重量为 3kg”	笔记本电脑的价格是“便宜”
若“CPU 速度快”	笔记本电脑的价格是“昂贵”

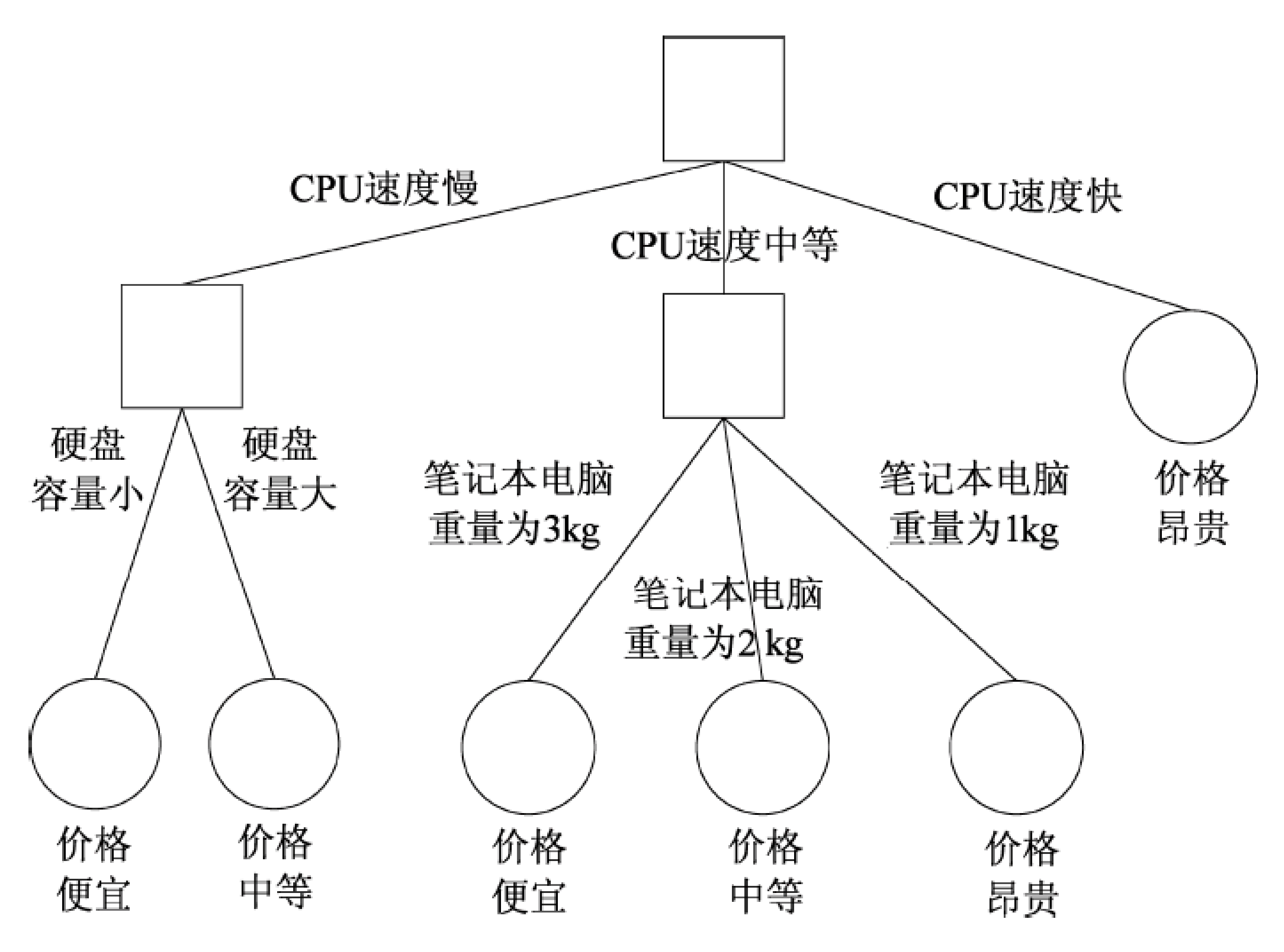


图 4.11　决策树树状架构图

重复选取属性为分支变量不仅产生多余的规则，也会造成决策树过于庞大而不容易解释，因此，适当地合并规则，可以使决策树的应用更具效率。例如，以下是通过某决策树所提取出的两条规则：

(1) IF "属性 U 小于 10、属性 V 为 i 或 j、属性 U 小于 5、属性 V 为 j"，THEN"被归纳于类别 A"。

(2) IF "属性 S 为 p、属性 T 大于 30、属性 W 为 l 或 m、属性 T 大于 40"，THEN"被归纳于类别 C"。

通过合并的方式可形成以下的新规则：

(1) IF "属性 U 小于 5、属性 V 为 j"，THEN"被归纳于类别 A"。

(2) IF "属性 S 为 p、属性 T 大于 40、属性 W 为 l 或 m"，THEN"被归纳于类别 C"。

决策树的分类结果容易受所选择的分支变量所影响，因此其他与分支变量相依性过高的因子，往往不易于决策树中被发现，以致错失找到真正关键因子的机会，进而减少数据判读的正确性。故在选取适当的目标变量后，必须检验各解释变量相依性高的其他变量，以分析不同的分支变量对目标变量的影响。

4.2　决策树的算法

决策树算法属于监督式学习法的一种，借由分类已知的事物来建立树状式结构，以从中归纳并提取规则，并进行未知样本的预测(Quinlan，1986)。决策树的层级架构，可以分析不同层级的变因对目标变量的影响，因此随着不同数据，采用不同的算法，得到的树状结构自然不同。根据目标变量的尺度又可将决策树分为分类树(classification tree)与回归树(regression tree)，最大的不同在于分类树的目标变量为类别形态；而回归树的目标变量则为连续形态。

目前常见的决策树算法包括：分类与回归树(classification and regression trees，CART)(Breiman *et al.*，1984)、卡方自动交互检测(Chi-squared automatic interaction

detection, CHAID)(Kass, 1980)、C4.5/C5.0(Quinlan, 1993, 1986)等一系列方法。其分支准则、分支方法与修剪方法的比较如表 4.8。

表 4.8 决策树算法比较

算 法		CART	C4.5/C5.0	CHAID
处理数据形态		离散、连续	离散、连续	离散
连续型数据分支方式		只分 2 支	不受限制	无法处理
分支准则	类别型相依变数	Gini 分散度指标	信息增益比	卡方检定
	连续型相依变数	方差缩减	方差缩减	卡方检定或 F 检定(需先转化为类别变量)
分支方法	类别型独立变量	二元分支	多元分支	多元分支
	连续型独立变量	二元分支	二元分支	多元分支(需先转化为类别变量)
修剪方法		成本复杂性修剪	基于错误的修剪	无

4.2.1 CART

CART 以 Gini 系数作为决定分支变量的准则,在每个分支节点进行数据分隔,并建立一个二分式的决策树,以决定最佳分支变量(Breiman *et al.*, 1984)。CART 的特色除了为二元分支算法外,并能处理类别型变量以及连续型变量的分类问题。

首先,给定一个节点 t,以 Gini 系数对分支变量进行二元分割,假设属性的分支水平为 s,t_{left} 与 t_{right} 分别为节点 t 的左、右子节点,并比较分支前后的纯度差异,如式(4.14):

$$\Delta Gini(s,t) = Gini(t) - [Gini(t_{\mathrm{left}}) + Gini(t_{\mathrm{right}})] \tag{4.14}$$

若 $\Delta Gini(s,t)>0$,表示子节点的纯度比其父节点的纯度高,则不考虑分支;若 $\Delta Gini(s,t)\leqslant 0$,则表示子节点的纯度比其父节点的纯度低,则作为该变量的候选分支水平,借由穷举搜索所有可能的分支水平,CART 算法在每一个可能的分支变量中会选择具有最大化纯度的分支水平作为候选分支依据,再经由比较所有候选分支变量中具有最大纯度作为节点的分支。

当利用训练数据表完成决策树的构建,CART 利用成本复杂性的修剪方法,以降低不必要的分支。

4.2.2 C4.5/C5.0

C4.5 以信息增益比作为决定分支变量的准则,且为多元分支决策树,C4.5 算法最常用于处理类别型数据,若遇连续型数据则需事先将其转化成类别变量。相较于其他分类算法的预测准确性、复杂度和训练时间,C4.5 决策树算法提供了较佳的准确性及数据解释能力。由于遗漏值会在建立决策树的过程中被忽略与取代,因此遗漏值不影响信息增益比的计算(Quinlan, 1993a)。C5.0(Quinlan, 1998a)是 C4.5 的进阶版,C5.0 增加了交互验证(cross-validation)与训练数据重复抽测的机制(boosting),与 C4.5 相比,不仅决策树的结果更准、计算速度更快,且需占用的内存资源也较少(Quinlan, 1998b)。

C5.0 的核心算法仍是以 C4.5 为主，以下主要说明 C4.5 的分支与修剪的过程。假设给定一个节点 t，C4.5 依据信息增益比的结果，穷举搜索所有可能的分支，从中选择具有最大信息增益比的分支变量作为该层决策树的分支变量，在每个节点计算其信息增益比是否大于 0，若有则继续分支长树，直到所有节点的信息增益比均小于 0 为止。

完成决策树的生长后，C4.5 的修剪方法是采用**基于错误的修剪**(**error-based pruning**)以比较一个父节点和其子节点的纯度。C4.5 采用悲观式估计分类错误率的概念，并直接用训练数据的结果估计分类错误率，假设在某一个叶节点有 N 笔数据，其中，共有 E 笔数据分类错误，可能的分类错误率应该大于 E/N，在此将 E 笔错误数据视为在 N 次实验中可能发生的结果，可能发生错误的次数为 $0,1,\cdots,E$，给定一信心水平(confidence level，CL)下，则可用二项分配(binomial distribution)估计该叶节点预测错误的概率，如式(4.15)所示：

$$CL = \sum_{x=0}^{E} C_x^N p^x (1-p)^{N-x} \tag{4.15}$$

其中，p 代表该叶节点错误分类的概率，N 为该节点 t 中的数据个数、x 代表该节点中可能被错误分类的数据数，E 代表该节点中被错误分类的最大数据数。若某一叶节点共有 6 笔数据($N=6$)，其中所有数据均属于同一类别($E=0$)，再给定一信心水平 0.25 下，由式(4.15)推导而得其分类错误的概率 p 为 0.206。

$$0.25 = \sum_{x=0}^{6} C_x^6 p^x (1-p)^{6-x}$$

该叶节点的错误分类成本为数据个数与估计分类错误率相乘后的结果，通过成本估算可知，若该节点的错误分类成本较其父节点的错误分类成本高时，则应该修剪属于叶节点的分支。

4.2.3 CHAID

CHAID 是 AID(automatic interaction detection)算法的延伸，根据卡方检定统计量的显著性检定，决定最佳分支属性，可以将属性划分为多个分支，为多元分支(multi-branch)决策树算法。CHAID 算法是以卡方检定的结果以决定分支属性，先由用户制订合并(merge)的门槛值 α_1 与分割(split)的门槛值 α_2、α_3，将每个属性值视为不同群组，若是顺序尺度数据，则需要将数据依序排列，每次两两检定为找相邻的两组作为可能的分支，借由列联表找出相对应的类别，采用两两分支检定的方式，计算出用于检定两分支是否有显著差异的 p-value 值，若该 p-value$>\alpha_1$，则合并此两分支成为新群组，并重复检查所有分支，直到所有分支两两检定的结果均为显著或已经仅剩两个分支。

接着，检查所有包含两种以上种类的分支节点，若节点内的检定结果为显著且 p-value $<\alpha_2$ 时，则将该节点中不同类别的样本划分至不同的分支节点。当属性数据发生遗漏值时，CHAID 会将所有遗漏值视为同一个群组，最后，有鉴于样本数会影响到分支检验，在 CHAID 中以 Bonferroni 调整 p-value 系数来做最终比较的依据(Kass，1980)。最后由所有 Bonferroni 调整 p-value$<\alpha_3$ 的属性中挑选最显著的属性作为分支节点，并将该节点中不同类别的样本区隔至不同的分支节点，否则即以此节点作为叶节点。在 CHAID 分支过程中，每个节点是基于选定的相依变量而分支，并以卡方检定作为分支准则以区隔分类属性的显著程度。因此，当所区隔的分支并无显著差异时，则合并为同一分支；反之，若具显著差异

时，则保留该分支并进行下一层的分支步骤。

CHAID 的最大限制在于数据特性必须为类别变量，倘若遇到连续型变量，则需将数据转换为类别型变量，或以高、中、低等类别属性来取代原有的数值变量。CHAID 算法与 CART、C4.5 的不同之处在于，后两者会采用事后修剪决策树的方式，而前者则于决策树的建立过程中，直接加入使决策树停止生长的机制。

4.3 决策树分类模型评估

决策树的分类模式随着不同算法而有不同的分类结果，可从两方面去评估其分类及预测表现：①以测试组数据的结果来客观评估较佳的决策树模型，例如分类错误率；②由于分类规则的提取随着问题而异，会因环境而造成规则解释的迥异，因此在客观评估后，通常均需由该领域专家根据问题背景选出最适合的决策树模型。

给定一组数据组 t_i 以及明确类别 C_j，由于该数据组 t_i 可能属于也有可能不属于该类别，假设有两个类别 Class 1 与 Class 2，例如良品与不良品。若分类模型预测结果与数据的实际类别一致，该结果为“真”(true)，若不一致，则该结果为“伪”(false)，也就是误判。常见的误判有两种：一为实际为 Class 1 但却判为 Class 2(false negative)；另一为实际为 Class 2 但却判为 Class 1(false positive)。依据预测结果与数据的实际类别，共有四种组合，二元类别的分类结果可产生一个混乱矩阵(confusion matrix)，见表 4.9。

表 4.9 二元问题的混乱矩阵

预测类别 / 实际类别	Class 1	Class 2
Class 1	TP(true positive)	FN(false negative)
Class 2	FP(false positive)	TN(true negative)

True Positive (TP)，预测为 Class 1 且实际为 Class 1；

False Positive (FP)，预测为 Class 1 但实际为 Class 2；

True Negative (TN)，预测为 Class 2 且实际为 Class 2；

False Negative (FN)，预测为 Class 2 但实际为 Class 1。

根据上述分类结果，可计算出**正确率(accuracy)**或**分类错误率(misclassification error rate)**如式(4.16)与式(4.17)所示：

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{4.16}$$

$$\text{Error rate} = 1 - \text{Accuracy} = \frac{\text{FP}+\text{FN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{4.17}$$

评估一个分类模型对两类别的分类或预测能力包含两个指标：一个是对欲辨识类别的**敏感度(sensitivity)**，亦即当该类别确实正确被预测的比率；**准确度(specificity)**则为另一个类别且确实被划分为另一个类别的比率。

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4.18}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.19}$$

分类模型的好坏除了以正确率高低来决定外,当某一类别的比率相对少,而该类别比另一个类别更受到重视,也就是类别的重要性可能不同,例如工程师想了解不良品发生的原因,然而,生产线的良品与不良品的比率往往相当不一致,此时仅用正确率可能会偏向都找到一堆指向良品的结果,而对工程师而言,更需要的是有关不良品的信息。

准确率(precision)与召回率(recall)也是常用的评估指标,如式(4.20)与式(4.21)。准确率指的是所预测的类别中,有多少比率的数据刚好属于该类别,准确率越高,表示该类别误判的比率越低。召回率则表示实际上为某类别的数据中,同时被判断为该类别的比率,其中召回率与敏感度的计算结果相同。

$$p = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.20}$$

$$r = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.21}$$

一般而言,准确率与召回率一个变大,另一个就会变小,因此,两个指标可合并成为一 F_1 综合性指标,如式(4.22)所示,F_1 值越高,表示该分类模型的准确率与召回率亦越高。

$$F_1 = \frac{2}{\frac{1}{r} + \frac{1}{p}} = \frac{2rp}{r + p} \tag{4.22}$$

实际上,当分类结果的判断会根据不同的门槛值而有所差异,也就是改变门槛值将增加或减少判断为 Class 1 的结果,进而影响敏感度与准确度时,可改用 ROC 曲线(receiver operating characteristic curve),如图 4.12 为例,其中,TP rate 为纵轴,FP rate 为横轴。TP rate 是描述当数据属于类别 C_1 时,被正确判断的概率;而 FP rate 则是当数据不属于 C_1 时,被误判概率。一般而言,TP rate 为越大越好,而 FP rate 为越小越好。因为准确度=1-FP rate,当敏感度(TP rate)增加时,准确度也会减少,也就是 FP rate 会增加,因此,ROC 曲线可作为衡量不同 FP rate 下 TP rate 的变化。

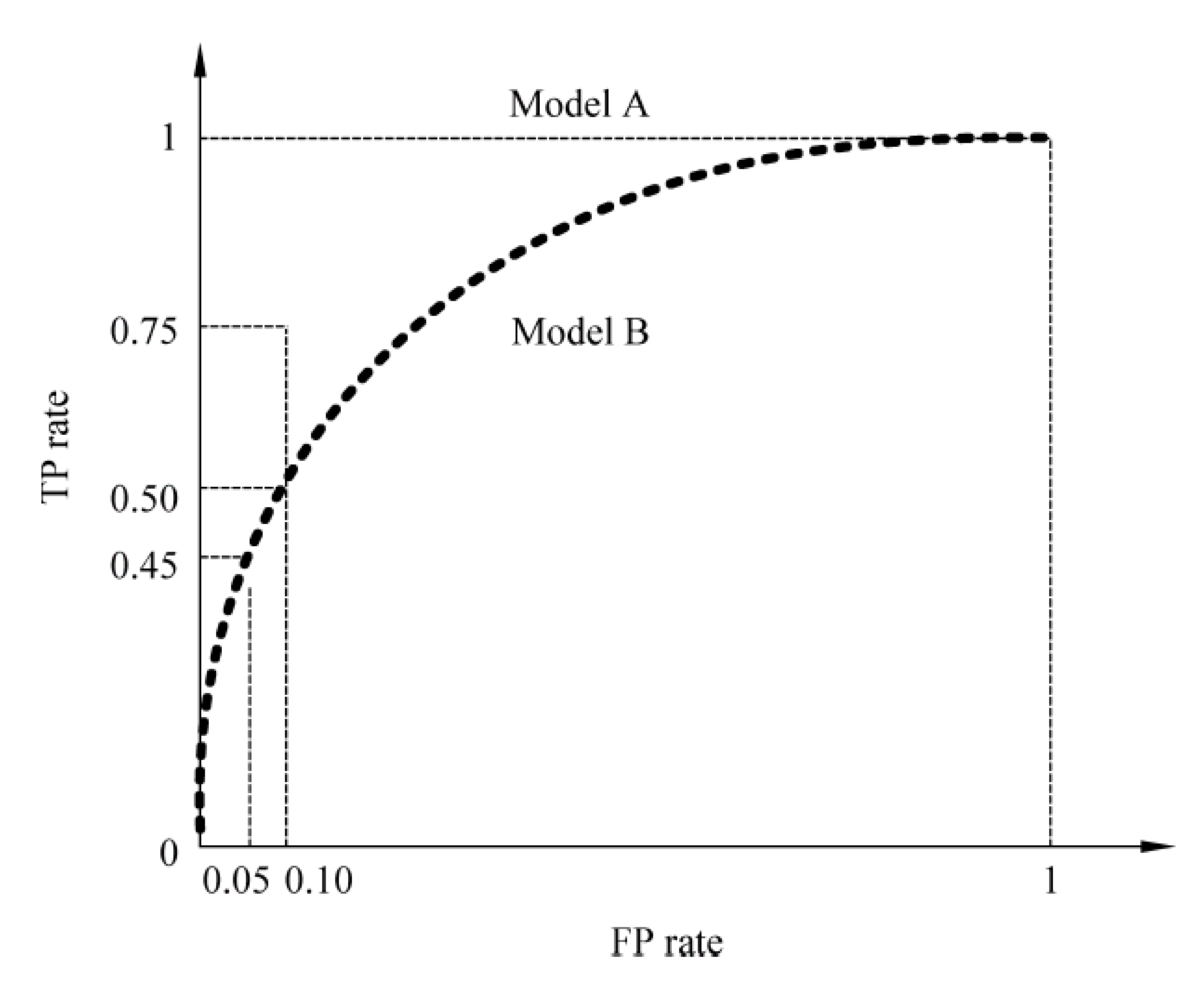


图 4.12　ROC 曲线

图 4.12 为两种方法 Model A 与 Model B 在不同分类判断门槛值下的 ROC 曲线，当期望 FP rate 小于 0.05 时，Model B 的分类结果较 Model A 好；如 FP rate 大于 0.05 时，则 Model B 的分类结果较 Model A 差；若可将 FP rate 放大至 0.10，可发现 Model A 的 TP rate 提升至 0.75，而 Model B 的 TP rate 则仅提升至 0.50。由此可知在不同的 FP rate 下，不同分类模式结果的比较差异。

一般而言，FP rate 的结果会根据分类门槛值的不同而有所变化，此时也可根据 ROC 曲线下的面积大小作为选择最佳分类结果模式。若 ROC 曲线下的面积越大，表示模式分类效果越好；反之，若该模型的分类能力不佳，其面积会越接近 0.5。

4.4　R 语言与决策树分析

本节采用美国国家糖尿病、消化与肾脏疾病研究所(US National Institute of Diabetes and Digestive and Kidney Diseases)对超过 21 岁的皮马族印第安人(Pima Indian)女性所做的糖尿病检测数据(Ripley，1996；Smith *et al.*，1988)，借以说明如何通过 R 语言使用 CART、C5.0 与 CHAID 三种决策树算法，分析哪些属性能帮助判断民众是否会得糖尿病。本组原始数据共包含 768 笔观测值以及 8 个属性，去除遗漏值后共剩下 532 笔完整数据，各属性尺度与属性值区间整理如表 4.10。

表 4.10　糖尿病检测数据

编号	属性名称	属 性 说 明	数据尺度	属　性　值
1	npreg	怀孕次数	连续	[0，17]
2	glu	葡萄糖浓度	连续	[56，199]
3	bp	血压	连续	[24，110]
4	skin	三头肌皮褶厚度	连续	[7，99]
5	bmi	身体质量指数	连续	[18.2，67.1]
6	ped	糖尿病家族病因指数	连续	[0.085，2.42]
7	age	年龄	连续	[21，81]
8	type	是否罹患糖尿病	类别	Yes，No

4.4.1　CART 决策树分析

扩充套件 **MASS**(Venables & Ripley，2002)中已将此数据集分为训练数据集(200 笔观测值)与测试数据集(332 笔观测值)。在呼叫内建的数据集后，便可利用扩充套件 **rpart**(Therneau *et al.*，2014)进行 CART 决策树构建。在此，先以不修剪的方式进行 CART 决策树的构建，故将函数中的复杂系数 cp 设定为 0。

```
library(MASS)
library(rpart)
data("Pima.tr")
summary(Pima.tr)
```

```
set.seed(1111)                                                    #设定随机数种子
cart=rpart(type~.,Pima.tr,control=rpart.control(cp=0))           #训练 CART 模型
summary(cart)
par(xpd=TRUE);plot(cart);text(cart)
```

图 4.13 为使用训练数据所构建而成的决策树模型，在完全未修剪的设定下共有 8 个叶节点以描述预测结果，所分支的属性包含葡萄糖浓度(glu)、年龄(age)、血压(bp)、糖尿病家族病因指数(ped)以及身体质量指数(bmi)等 5 项。

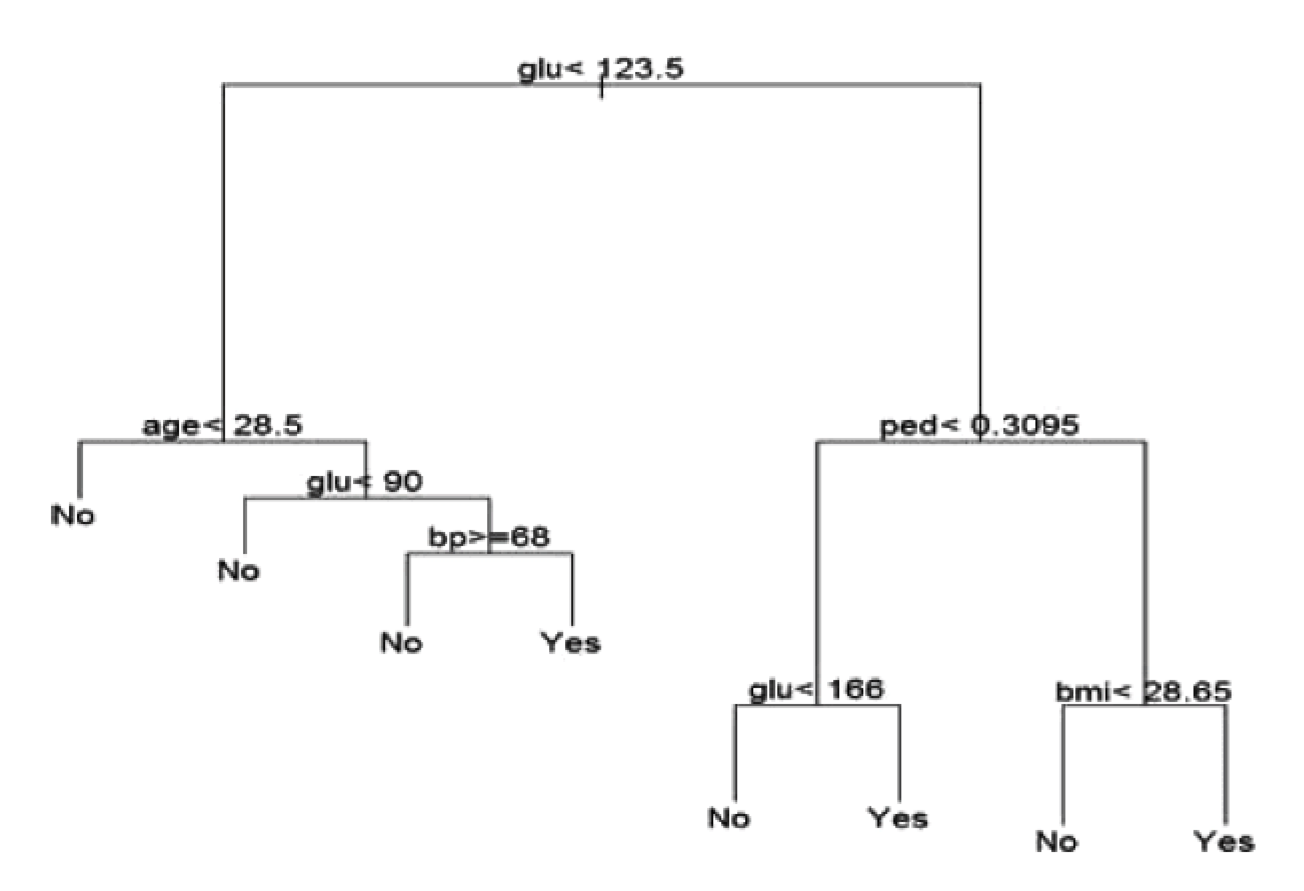
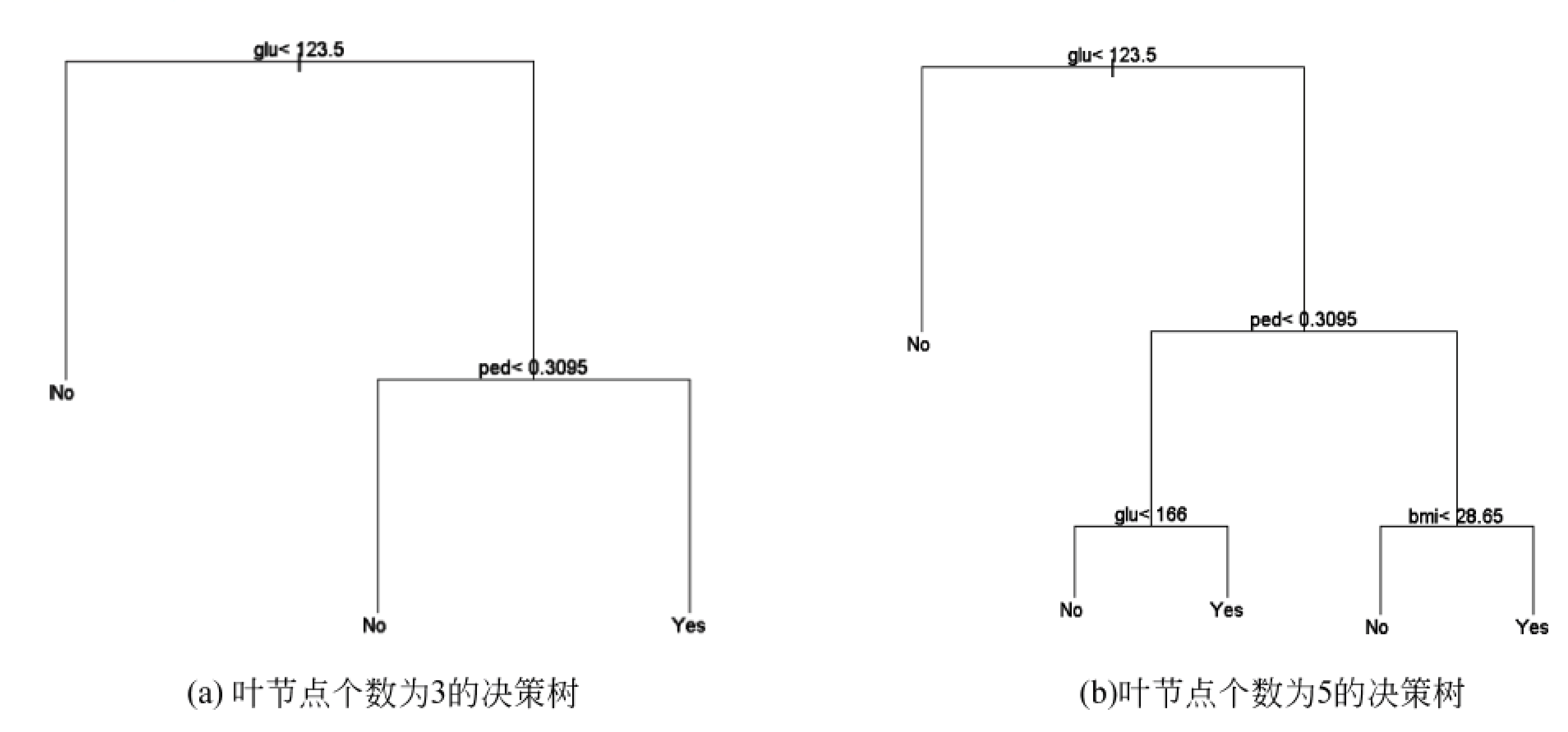


图 4.13　CART 未修剪决策树(叶节点个数为 8)

CART 是根据基于错误的修剪进行决策树的修剪，越小的复杂系数 cp(即为 α)代表叶节点个数越多，虽然对训练数据的解释力越高却也容易落入过度配适，失去对新数据的预测能力。在 R 语言中可以根据以下语法对已建好的 CART 模型给定复杂度参数进行修剪，并重新绘制决策树树形图。

```
cart_prune=prune(cart,cp=0.03) #cp=0.1=>叶节点数=3; cp=0.03=>叶节点数=5
par(xpd=TRUE);plot(cart_prune);text(cart_prune)
```

图 4.14 (a)与图 4.14 (b)分别是 CART 决策树进行叶节点数为 3 与 5 的修剪后所得到的图形，从中可看出所用到的属性已只剩葡萄糖浓度(glu)、糖尿病家族病因指数(ped)以及身体质量指数(bmi)3 项。



(a) 叶节点个数为3的决策树　　(b)叶节点个数为5的决策树

图 4.14　CART 修剪后的决策树

此外,也利用测试数据集检验训练数据所产生的 CART 决策树模型。

```
pre=predict(cart,Pima.te,type="class")
confusion_matrix=table(Type=Pima.te$type,Predict=pre)          #建立预测交叉矩阵
confusion_matrix
accuracy=sum(diag(confusion_matrix))/sum(confusion_matrix)     #计算正确率
accuracy
#将第一行指令的 cart 置换成 cart_prune 便可利用修树后的模型进行预测
#将第一与第二行的 Pima.te 置换成 Pima.tr 便可计算模型对训练数据的正确率
```

表 4.11 列出三种 CART 模型在训练数据集与测试数据集上的正确率。从中亦同样可看出,叶节点个数越多,虽然训练数据的正确率越高,但测试数据的正确率在叶节点个数为 5 时最高,显示复杂的决策树反而失去其预测能力。

假设修剪后的叶节点个数为 3 或 5。选择叶节点个数为 3 的修剪原因在于此时错误率已达稳定,虽未达最低但规则较不复杂且容易解释;选择叶节点个数为 5 的修剪原因是希望获得较高的预测准确率。

表 4.11 CART 决策树模型正确率比较

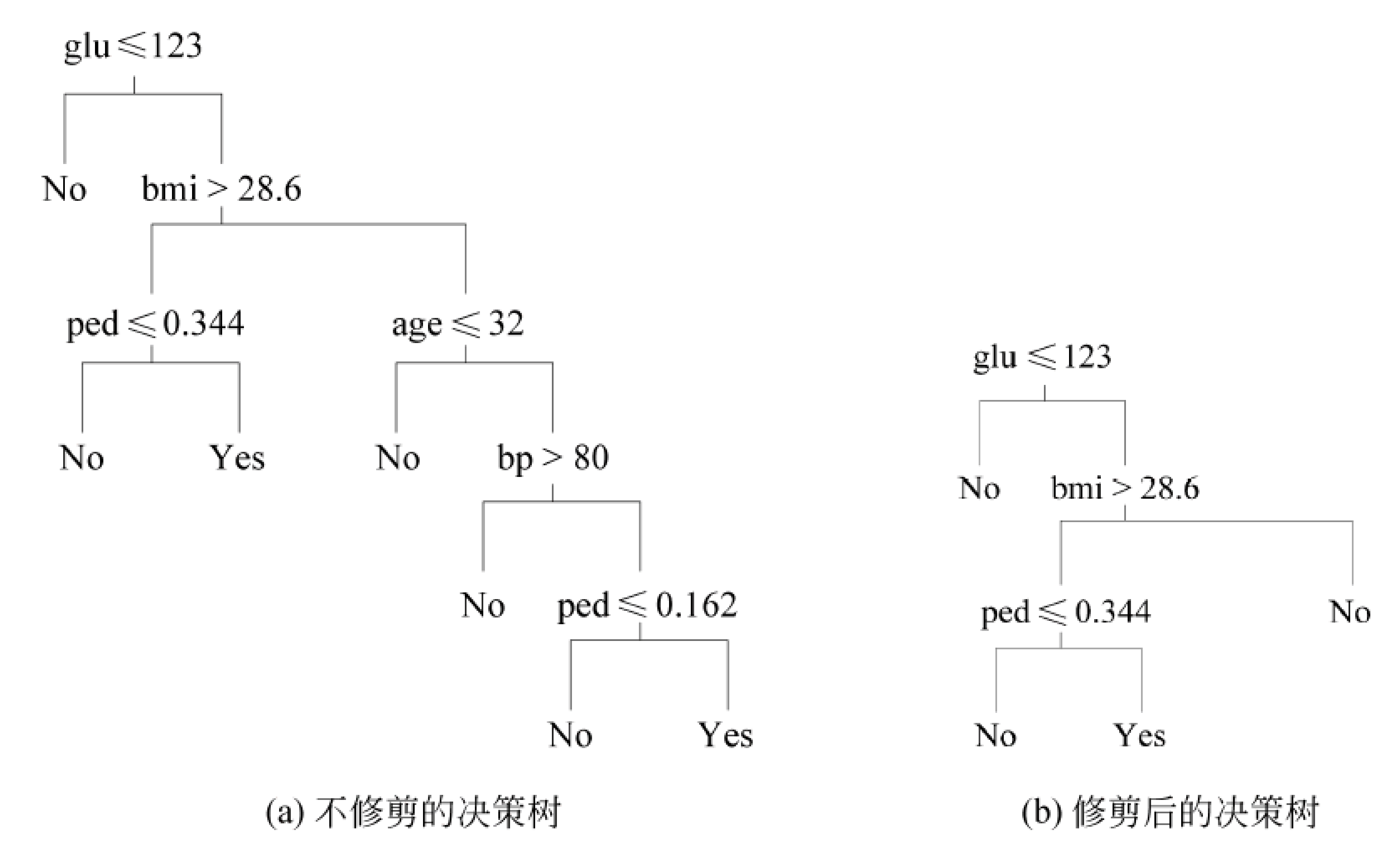
	叶节点数为 8(不修剪)	叶节点数为 5	叶节点数为 3
训练数据	0.850	0.835	0.790
测试数据	0.732	0.756	0.729

4.4.2 C5.0决策树分析

以下利用扩充套件 **C5.0**(Kuhn *et al*.,2014)进行 C5.0 决策树构建。在此,先以不修剪分支下构建决策树,将函数中的 *noGlobalPruning* 参数设定为 T。

```
library(C50)
library(MASS)
data("Pima.tr")
C50_tree=C5.0(type~.,Pima.tr,control=C5.0Control(noGlobalPruning=T))
summary(C50_tree)
#若将 noGlobalPruning 参数设定为 F 则会进行修树功能
```

图 4.15(a)与图 4.15(b)分别为未进行修树与修树后的树形图。未进行修树的结果共产生 7 个叶节点,所使用的属性包含葡萄糖浓度(glu)、年龄(age)、血压(bp)、糖尿病家族病因指数(ped)以及身体质量指数(bmi)等 5 项;而经过修剪后的决策树则剩下 4 个叶节点,使用的属性包含葡萄糖浓度(glu)、糖尿病家族病因指数(ped)以及身体质量指数(bmi)等 3 项。此结果与 CART 决策树相同,唯属性的出现顺序以及分支门槛值不同。另外,表 4.12 则呈现修剪前后对于决策树模型的正确率。虽然修剪前后对于测试数据正确率并无差异,但在模型解释上仍以修剪后的结果较容易被解释。



(a) 不修剪的决策树　　(b) 修剪后的决策树

图 4.15　C5.0 决策树

表 4.12　CART 决策树模型正确率比较

	叶节点数为 7(不修剪)	叶节点数为 4
训练数据	0.840	0.815
测试数据	0.735	0.735

4.4.3　CHAID 决策树分析

利用扩充套件 **CHAID**(The FoRt Student Project Team,2013)构建 CHAID 决策树。CHAID 算法是以卡方检定为基础进作为分支准则,因此不用考虑事后修剪,由于 CHAID 算法却仅能处理类别型的属性,因此,必须先将数据中连续值的属性进行离散化。以下是通过分箱法将所有连续型属性进行 3 等份分割的程序,7 个离散化后的属性形成顺序尺度,其分割水平如表 4.13 所示。

```
library(CHAID)
library(MASS)
data("Pima.tr")
data("Pima.te")
Pima=rbind(Pima.tr,Pima.te)
level_name={}
for (i in 1:7) {
  Pima[,i]=cut(Pima[,i],breaks=3,ordered_result=T,include.lowest=T)
  level_name<-rbind(level_name,levels(Pima[,i]))
}
level_name=data.frame(level_name)
row.names(level_name)=colnames(Pima)[1:7]
colnames(level_name)=paste("L",1:3,sep="")
level_name
```

表 4.13 离散化后属性

属性	水平 1	水平 2	水平 3
npreg	[−0.02,5.66]	(5.66,11.3]	(11.3,17]
glu	[55.9,104]	(104,151]	(151,199]
bp	[23.9,52.6]	(52.6,81.4]	(81.4,110]
skin	[6.91,37.6]	(37.6,68.4]	(68.4,99.1]
bmi	[18.2,34.5]	(34.5,50.8]	(50.8,67.1]
ped	[0.08,0.86]	(0.86,1.64]	(1.64,2.42]
age	[20.9,41]	(41,61]	(61,81.1]

接着，以预设检定显著水平为 0.05 进行 CHAID 之决策树构建，并以前 200 笔数据为训练集，后 332 笔数据为测试及验证效度。

```
Pima.tr= Pima[1:200,]
Pima.te= Pima[201:nrow(Pima),]
set.seed(1111)
CHAID_tree= chaid(type~ .,Pima.tr)
CHAID_tree
plot(CHAID_tree)
```

图 4.16 为 CHAID 决策树的模型，共有 5 个叶节点，使用的属性包含葡萄糖浓度(glu)、年龄(age)与糖尿病家族病因指数(ped)等 3 项。

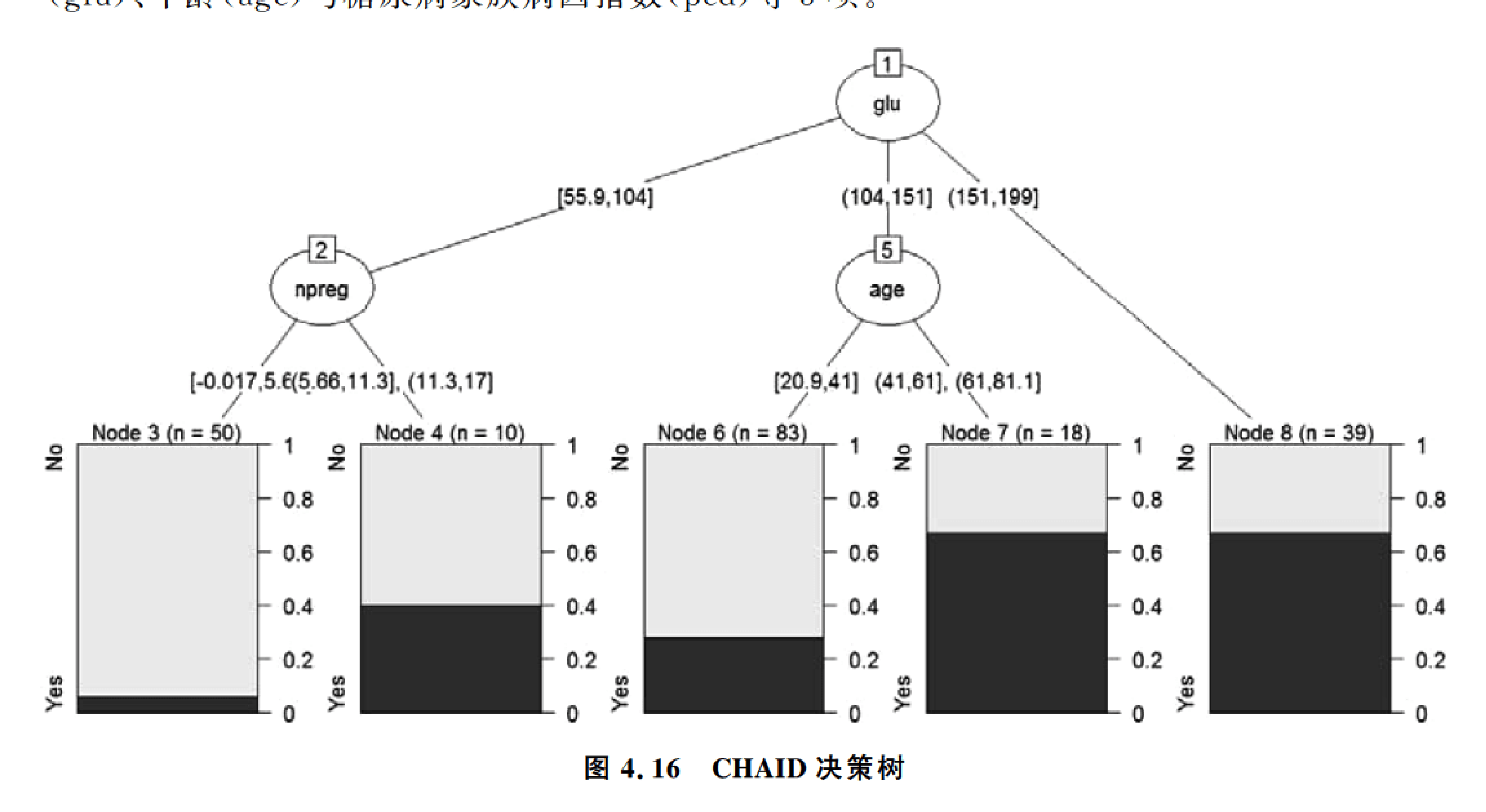


图 4.16 CHAID 决策树

CHAID 决策树模型于训练数据与测试数据的正确率分别为 0.755 与 0.789。如表 4.14，CHAID 的测试数据正确率高于 CART 与 C5.0 的结果。

表 4.14　三种决策树算法结果

比 较 项 目	CART	C5.0	CHAID
训练数据正确率	0.835	0.815	0.755
测试数据正确率	0.756	0.735	0.789
叶节点数	5	4	5
深度	3	3	2
使用属性	葡萄糖浓度、糖尿病家族病因指数、身体质量指数	葡萄糖浓度、糖尿病家族病因指数、身体质量指数	葡萄糖浓度、年龄、糖尿病家族病因指数

4.5　应用实例——建构 cDNA 生物芯片的数据挖掘模式

4.5.1　案例背景

一片生物芯片可同时解析出上千种基因表现,庞大的数据,若未经进一步的数据处理和分析,将难以从中发现致病的基因。目前一片芯片的价格高达 500 美元,因此受测者的样本数往往远小于实验变量个数,使得数据搜集不易,不仅增加生物芯片分析的困难度,更遑论检查各个基因彼此之间的交互作用。本案例(简祯富,林国胜,2006)针对生物芯片上 cDNA 数据应用决策树分析方法,搜索出基因在正常人与病人中不同的表征,以及借由了解基因与致病因子之间的关联,结合生物医学研究者其领域知识发展,发掘出有意义的信息,以提供医学研究者针对特定的疾病或症状下判断的依据。

本案例选用斯坦福大学的生物芯片数据库(Stanford Microarray Database, SMD)(http://smd.princeton.edu/)中乳癌实验芯片 cDNA 数据进行研究,各芯片约包含 45 696 个基因(探针点)与病人、非病人各一位样本,反应后所得的表现值,总计 64 笔芯片数据,原始数据内含编号 18 196 芯片数据,每列为各个不同基因,每栏表示各个基因不同表现值,包含基因名称、坐标、基因强度表现等,共 128 笔样本,病人与非病人各半,如 spot 为探针流水编号、Accession 为基因名称,而 Ch1/Ch2 Net 的数值为各基因相对应的正常人/病人基因强度表现。

4.5.2　数据准备

首先整理各芯片数据以去除冗余及不需要的名目字段,如 spot、gene name、gene symbol、gene ID,仅保留 Accession No.(即唯一且统一的完整基因编码),再去除不需要及无效的数值字段,如 Accession No. 名称遗失,以及个别 Accession No. 遗漏值过多者(20%遗漏值,本数据为遗漏值超过 25 个)。针对无法控制或判别的潜在变异,即基因 i 非某特定疾病的显著基因,若某些正常人基因 i 表现异常的离群值,为避免误判予以删除。数据准备后,共计 41 681 个基因(列)与 128 笔样本(栏)。接着将整理后的数据集采用随机重复抽样的方式,并借用交互验证(cross-validation),训练集数据与测试集数据为 80%与 20%,分别区分五次个别的训练集与测试集,各包含 100 笔与 28 笔样本。重复抽样 n 次以计算平均正确率,各次训练数据集主要用以构建生物芯片数据的决策树模式与规则,测试集数据则用以

衡量模式的效度。

4.5.3 生物芯片数据的决策树构建

将处理过的病人与非病人各 64 笔数据所汇整的数据表，任意成对挑选出各 50 笔，共 100 笔作为训练数据，剩余各 14 笔，共 28 笔作为最后验证用数据，并重复抽样五次。显著规则筛选后分别得到 11 104、12 829、13 219、12 770、13 745 个较显著基因。接着将筛选出的较显著基因当做乳癌决策树的分类属性，经由五次重复抽样实验后分别得到 12、14、18、14、16 个分支，综合其解释率达到 90%以上的决策树，共得到 21 个分支(如图 4.17)，汇整影响乳癌的基因及其 IF-THEN 规则及其分支正确率(判定为乳癌病人之分支)、平均正确率及模式解释力等，其中分支解释率以分数形式表示能更清楚显示分支情形，括号内的数字为该决策规则在五次重复抽样分析中出现的次数，如 50/50(5)为此规则 50 人为判定乳癌患者，实际患者亦为 50 人，此规则在五次重复抽样中出现五次；整体正确率为计算所有正确判别的比率；平均模式解释力为单一规则在各次模式解释力的平均表现，详见表 4.15。

Rule 1：IF (AA777396＜1000) THEN patients(若基因 AA777396＜1000，则判定为患有乳癌)；

Rule 2：IF (AA985123＜1000) THEN patients(若基因 AA985123＜1000，则判定为患有乳癌)；

Rule 3：IF (AA961402＜2000) THEN patients(若基因 AA961402＜2000，则判定为患有乳癌)。

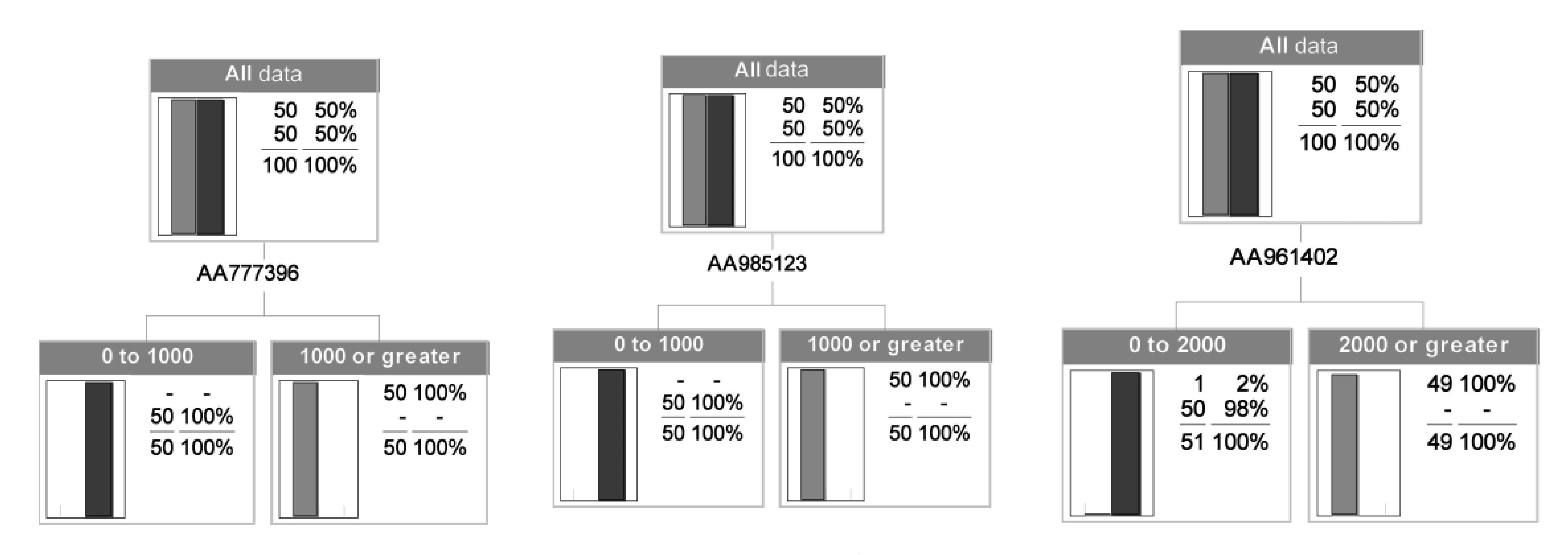


图 4.17 部分决策树规则

表 4.15 决策树规则整理表

项次	决 策 规 则	分支正确分类率(次数)	平均整体正确率	平均模式解释力
1	IF (AA777396＜1000) THEN patients	50/50(5)	1.00	100%
2	IF (AA985123＜1000) THEN patients	50/50(5)	1.00	100%
3	IF (R95691＜5000) THEN patients	49/49(4) 50/50(1)	0.99	97%
4	IF (H79533＜5000) THEN patients	50/51(5)	0.99	96%

续表

项次	决 策 规 则	分支正确分类率(次数)	平均整体正确率	平均模式解释力
5	IF (T53121<2000) THEN patients	50/51(5)	0.99	96%
6	IF (AA961402<2000) THEN patients	50/51(1)	0.99	96%
7	IF (AA938940<1000) THEN patients	50/51(1)	0.99	96%
8	IF (AA913206<5000) THEN patients	49/49(2) 49/50(2)	0.99	94%
9	IF (AA701996<2000) THEN patients	48/48(4) 49/49(1)	0.98	93%
10	IF (AI380522<2000) THEN patients	48/48(4) 49/49(1)	0.98	93%
11	IF (T98611>5000) THEN patients	48/48(5)	0.98	92%
12	IF (AI679372>5000) THEN patients	48/48(1)	0.98	92%
13	IF (AA233079<5000) THEN patients	50/52(3)	0.98	92%
14	IF (W56522<2000) THEN patients	50/52(5)	0.98	92%
15	IF (AI001134<2000) THEN patients	50/52(4) 48/48(1)	0.98	92%
16	IF (AI375135<4000) THEN patients	50/52(2)	0.98	92%
17	IF (AI923787>2000) THEN patients	49/50(4) 48/48(1)	0.98(5)	92%
18	IF (H52245>2000) THEN patients	48/48(2)	0.98(2)	92%
19	IF (W01204<2000) THEN patients	50/52(3)	0.98(3)	92%
20	IF (AA486362>2000) THEN patients	49/50(1)	0.98	92%
21	IF (H12338>1000) THEN patients	49/50(1)	0.98	92%

将各次重复抽样所剩余的 28 笔数据当作测试集进行模式规则验证，在医学上伪阴性(FN rate)(即实际上有病者未被检验出得病)较伪阳性(FP rate)显得重要，根据生物芯片与生物信息领域知识，若伪阴性(FN rate)高于 10%则该规则予以删除，若伪阳性(FP rate)高于 20%时则删除。将测试集数据分别带入各次分析中所挖掘出的决策规则中，在经过五次验证数据测试之后删除项次 5、6、7、8，IF (T53121<2000) THEN patients、IF (AA961402<2000) THEN patients、IF (AA938940<1000) THEN patients、IF (AA913206<5000) THEN patients 等四条规则，五次验证结果正确率均达 97%以上。

4.5.4　规则解释与评估

由于医学研究往往牵涉患者的健康、生命安全，研究模式的解释能力需以更严格的标准衡量，因此本研究选取模式解释能力 90%以上的 21 个基因为医疗检测参考因子并建立其个别决策规则，如当基因 AA777396 检测值小于 1000 时，则判定为患有乳癌，大于 1000 时则为正常人；在使用测试集进行验证分析时，采取伪阴性(FN rate)若高于 10%时则删除该

规则，伪阳性(FP rate)高于20%时予以删除，共删除三条决策规则(各规则信度效度如表4.15所示)。本案例所建立的决策规则，系纯以芯片数据进行分析，后续可整合相关病历数据，以更深入探讨病人基因表现值与不同病人特性之关系，如年龄、性别等。

本案例所提出的生物芯片决策树分析提供一个有效的方法，由乳癌实验芯片cDNA数据的分析结果验证其可行性。随着生命科学的知识及技术的快速发展，生物信息发现所累积的大量数据难以仅依靠传统统计技术，从生物芯片探索基因的影响为例，生物芯片一次就能记录成千上万个基因表现的样型，却因现实环境仅有少数样本的问题，在传统统计假设上，即因自由度的关系而无法进行实验设计，亦难以处理复杂交互作用情形下的分析。

4.6 结论

决策树在数据挖掘中常扮演监督式特征提取与描述的角色，经常用于解决分类的问题，并作探索与预测之用(Berry & Linoff，1997)，其预测技术乃是依据某一特定对象属性，观察其过去的行为或历史数据，借以估计未来的预测值。决策树在其分支节点会计算所选择区隔变量的显著程度。若是一次选择一个变量进行切割，则为单变量决策树(或为一般所称的标准决策树算法)；若选择的是变量的线性组合，则称为多变量决策树。

事实上，决策树并非唯一的分类工具，其他如人工神经网络等也可应付复杂且难以区隔的类别，但其模型或数学式相对难以解释。决策树分析对于高维度的数据也可快速学习，并构建层级式的树状结构，而挖掘所得的结果也可转换为一系列容易了解的IF-THEN规则，因此适合用来挖掘未知的知识或样型。

问题与讨论

1. CART、C4.5与CHAID为目前构建决策树较常使用的算法，请比较三者的优缺点与适用状况。

2. 下表整理了20位受检者的基本数据。假设有兴趣的目标变量为受检者是否驼背，请回答下列各问题。

(1) 请计算目标变量“驼背”分布于种类“是”与“否”所带来的信息总和。

(2) 请分别计算目标变量“驼背”经过水平“年龄(>50)”、“年龄(≤50)”、“身高(>175)”、“身高(≤175)”、“性别(男)”、“性别(女)”修正后的信息量。

(3) 请分别计算属性“年龄”、“身高”、“性别”对于目标变量“驼背”所带来的信息总量与信息贡献度。

(4) 请计算目标变量“驼背”的不纯度总和。

(5) 请分别计算目标变量“驼背”在各属性水平“年龄(>50)”、“年龄(≤50)”、“身高(>175)”、“身高(≤175)”、“性别(男)”、“性别(女)”下的不纯度。

(6) 请分别计算以属性“年龄”、“身高”、“性别”做分支，对于目标变量“驼背”的纯度所得。若以Gini指标作为决策树分支的准则，何者会优先列选为第一次分支的变数？

(7) 请分别计算属性“年龄”、“身高”、“性别”对于目标变量“驼背”的卡方统计量。若以卡方统计量作为决策树分支的准则，何者会优先列选为第一次分支的变数？

（8）请分别计算属性“年龄”、“身高”、“性别”对于目标变量“驼背”的信息增益比。若以信息增益比作为决策树分支的准则，何者会优先列选为第一次分支的变数？

心血管疾病数据表

编号	驼背	年龄（＞50）	身高（＞175）	性别
1	是	是	是	男
2	否	否	是	男
3	否	是	否	女
4	否	否	否	女
5	否	是	否	男
6	是	是	否	女
7	否	否	否	男
8	否	否	否	女
9	是	否	是	男
10	否	否	否	女
11	否	否	否	男
12	否	是	否	女
13	否	是	否	女
14	否	否	否	女
15	否	否	否	男
16	是	是	是	男
17	是	是	否	男
18	否	否	否	男
19	否	是	否	女
20	否	是	否	女

3. 某医院欲研究某心血管疾病的造成因子，分别收集了 5 个病患与 15 个正常人的年龄、血压与血型三项属性变量如下表所示。请根据数据回答下列问题：

（1）请问目标变量（观测体健康/生病状况）的信息总和为多少？

（2）请问在属性血压中，经过“偏低”、“正常”与“偏高”三种水平修正后，其信息总量分别为多少？

（3）请问三种属性对目标变量所带来的总信息量与信息贡献度分别为多少？在此例中，何种属性是用来预测观测体生病与否的最佳属性？

心血管疾病数据表

属性 1：年龄

年龄/岁	健康	生病	总和
0～25	4	1	5
26～40	7	2	9
41～	4	2	6
总和	15	5	20

属性 2：血压

血压	健康	生病	总和
偏低	2	0	2
正常	11	0	11
偏高	2	5	7
总和	15	5	20

属性 3：血型

血型	健康	生病	总和
O	7	2	9
A	3	2	5
B	3	0	3
AB	2	1	3
总和	15	5	20

4. 假设在制造过程中出现的异常是由某些因素造成的，请使用决策树找到可能的原因。并请利用表中的数据来计算下列数值：

(1) “产品是否有缺陷”的信息总和。

(2) 各属性解释“产品是否有缺陷”的纯度所得。

(3) 各属性的熵(entropy)。

(4) 各属性解释“产品是否有缺陷”信息增益比(gain ratio)。

(5) 各属性对“产品是否有缺陷”卡方统计量(Chi-square statistic)。

编号	站别 A	站别 B	站别 C	产品是否有缺陷
1	A01	B01	C03	N
2	A01	B01	C03	N
3	A02	B03	C01	Y
4	A03	B02	C02	Y
5	A03	B02	C03	Y

续表

编号	站别 A	站别 B	站别 C	产品是否有缺陷
6	A03	B01	C03	N
7	A02	B02	C01	Y
8	A01	B03	C02	N
9	A01	B02	C01	Y
10	A03	B03	C03	Y

5. 某项就业调查数据如下表所示，分别记录十个受访者的月收入、学历、产业别与性别等数据，请回答下列问题：

(1) 请问月收入的样本总方差为何？

(2) 请问分别以学历、产业别与性别分类后，月收入的总方差缩减程度分别为何？

(3) 假设分析者想要选择一项属性来解释各受访者的收入差异来源，何者为最佳的解释属性？

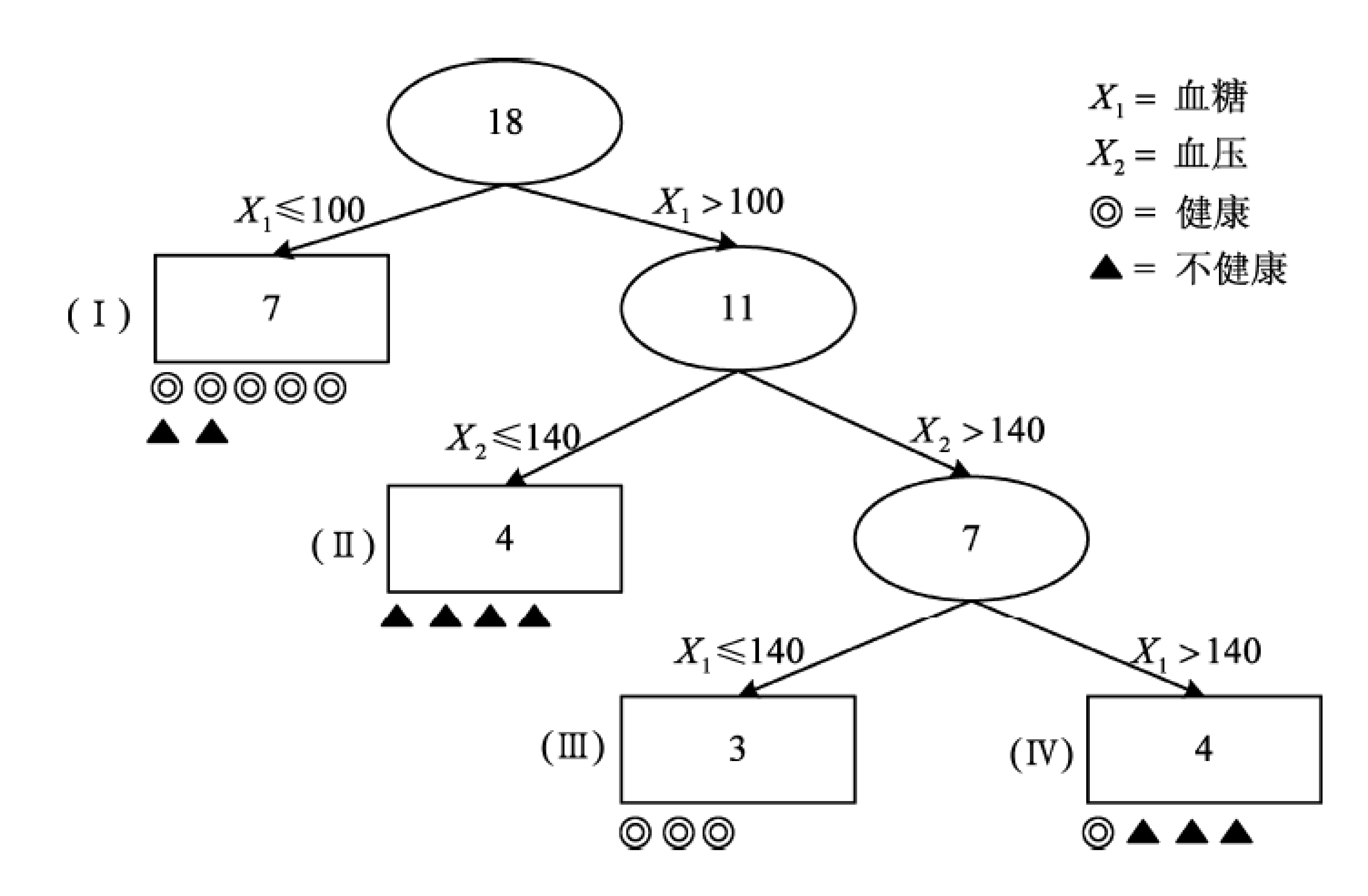
就业调查数据

月收入/千元	学历	产业别	性别
35	高中	A	男
42	大学	B	女
36	研究所	A	男
38	大学	B	男
22	高中	A	女
27	高中	C	男
53	大学	C	男
37	大学	C	女
42	研究所	C	女
71	研究所	B	男

6. 请根据以下决策树层级架构回答下列问题。

(1) 请由下列决策树中提取决策规则。

(2) 请计算决策规则的支持度、置信度、增益。



7. 试回答下列问题：

(1) 请根据图 1 画出决策树树形图，并计算每个叶节点的正确率。

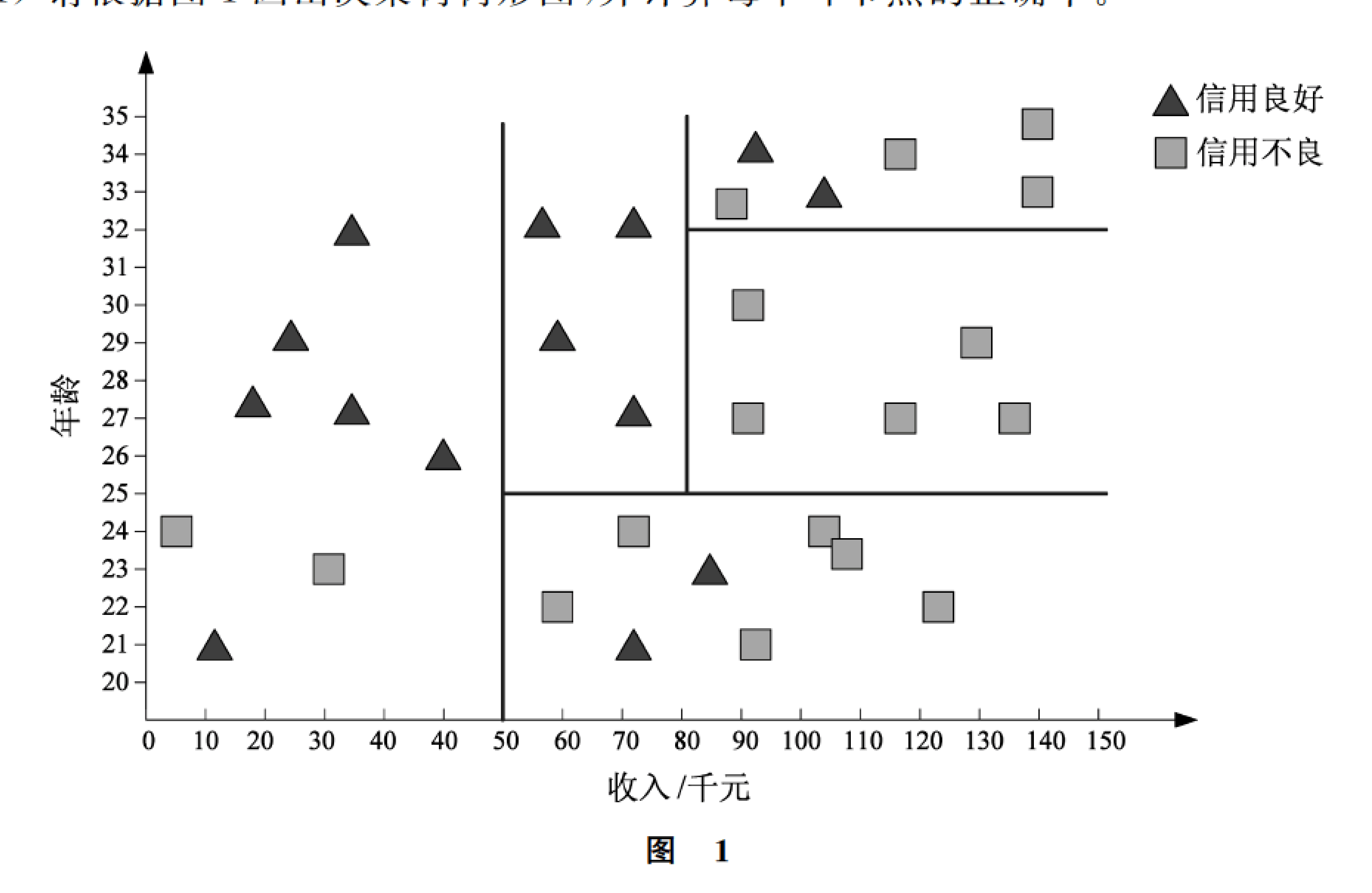


图 1

(2) 若另一决策树分析结果如图 2，请画出其决策树树形图，并分别以信息增益比和 Gini 两种指标，比较图 1 和图 2 所示的决策树的差异。

8. 下图为 20 个样本在连续属性 A 与 B 上的散布图，目标变量由符号●与▲表示其两种不同的类别。试回答下列问题：

(1) 假设分析者欲以二元分支的方式对此数据进行决策树的构建，其考虑了两种第一个分支的状况：(Ⅰ) $A>a$，$A<a$ 与(Ⅱ) $B>b_1$，$B<b_1$，请分别就 Gini 系数与卡方统计量为根据，说明何种分支方式较佳。

(2) 承上题，假设在选择(Ⅰ)的情况下，在分支的子节点中是否还存在较佳的分支方式？若有，其分支方式为何？

(3) 承题(1)，假设在选择(Ⅱ)的情况下，在分支的子节点中是否还存在较佳的分支方式？若有，其分支方式为何？

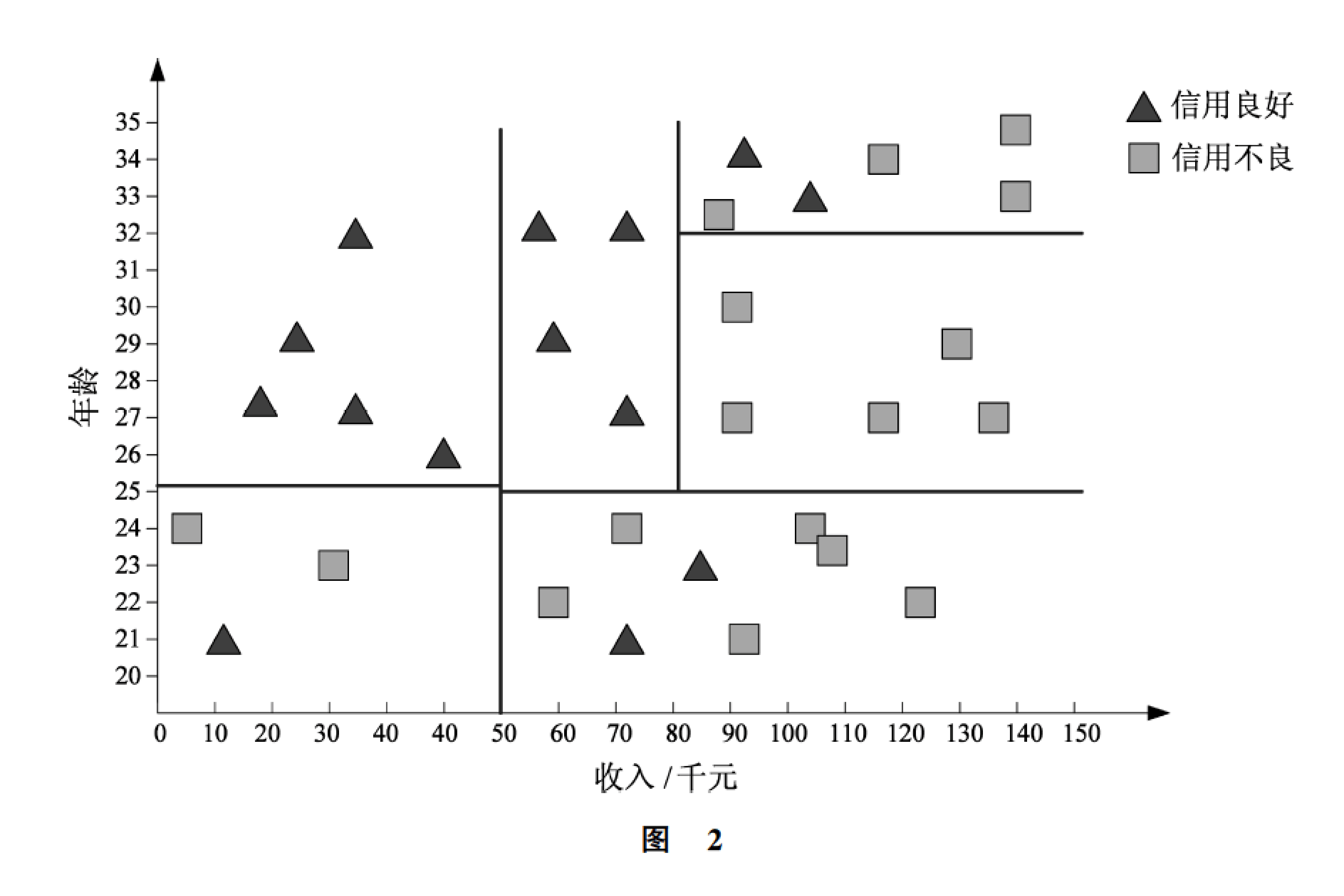
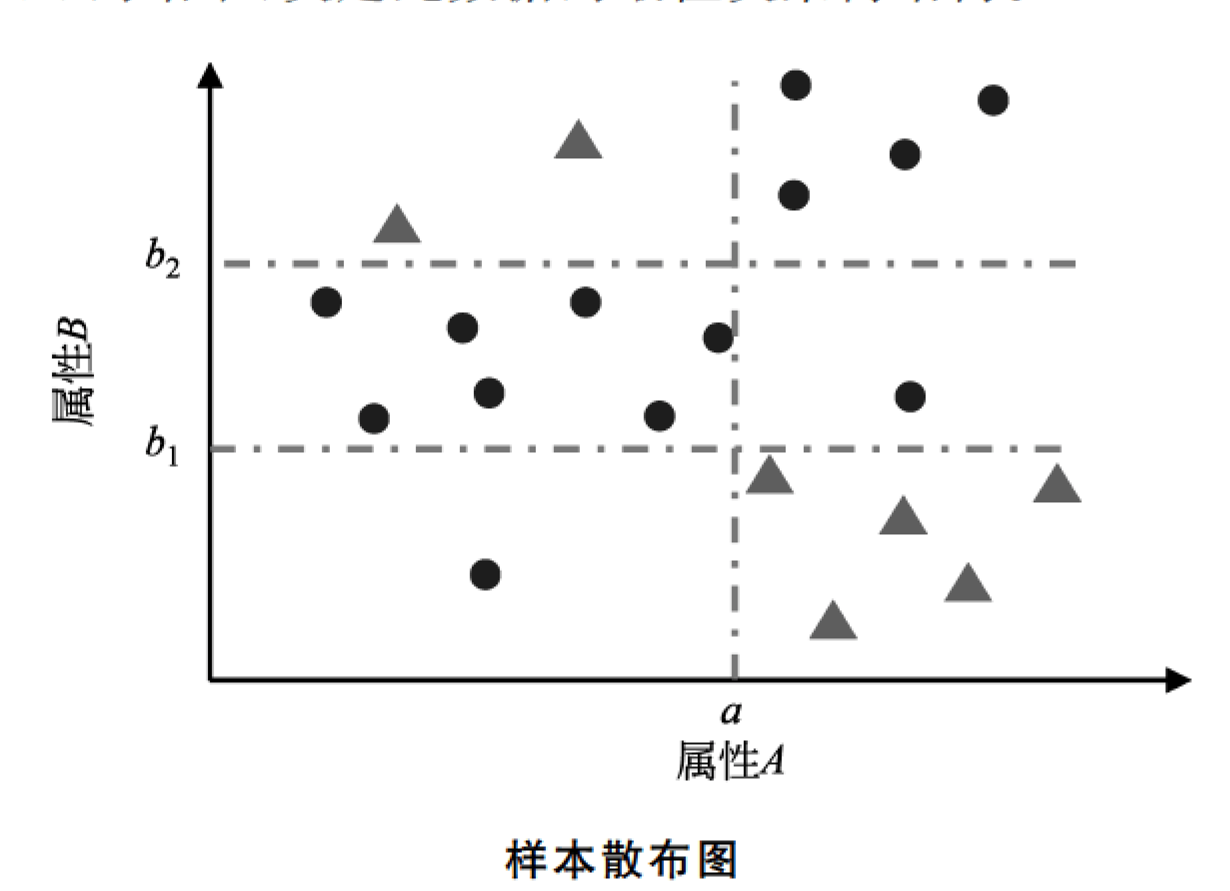


图　2

(4) 请就(1)～(3)的结果，决定此数据的最佳决策树结构。



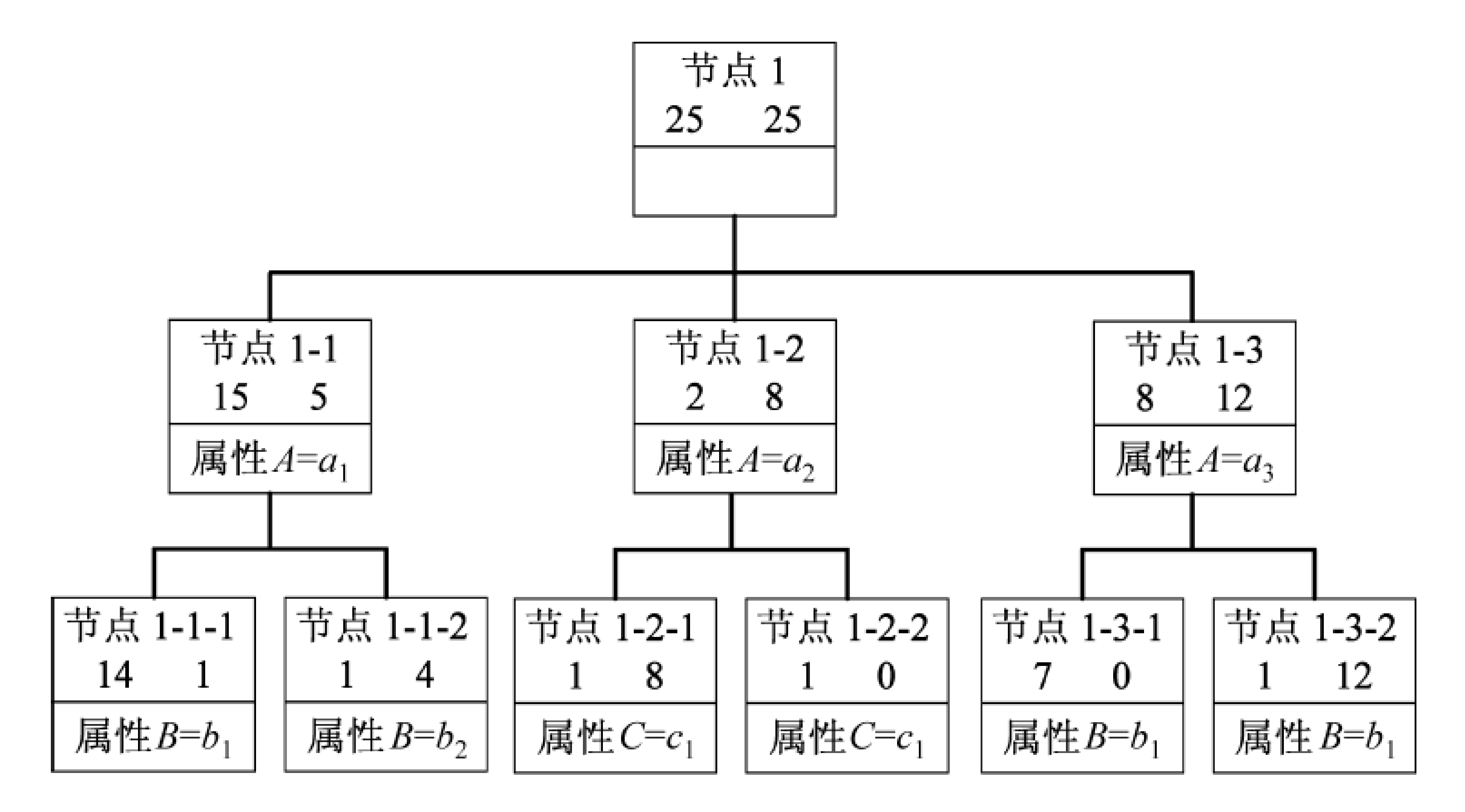
样本散布图

9. 假设某决策树在其节点 1 之下的分支如下图所示，请分别进行下列分析：

(1) 以最小成本复杂度修剪的方式修树(分别考虑 $\alpha=0.05, 0.1$)。

(2) 以最小错误修剪的方式修树。

(3) 假设节点 1 为此树的根节点，请根据以上结果进行规则提取。


节点 1
25 25
节点 1-1
15 5
属性 A=a1
节点 1-2
2 8
属性 A=a2
节点 1-3
8 12
属性 A=a3
节点 1-1-1
14 1
属性 B=b1
节点 1-1-2
1 4
属性 B=b2
节点 1-2-1
1 8
属性 C=c1
节点 1-2-2
1 0
属性 C=c1
节点 1-3-1
7 0
属性 B=b1
节点 1-3-2
1 12
属性 B=b1

决策树分支结构图

第 5 章

人工神经网络

人脑预估有超过 1000 亿个神经细胞（nerve cells），每个神经细胞借由许多突触（synapses）与其他神经细胞相连成一个非常复杂的神经网络，这些神经细胞以平行交织的方式来分析大量数据。当受到刺激，信号便经由神经细胞依续传递到大脑，大脑会下达指令做出相关反应，经由反复训练后，则会将此过程记忆于脑中。

神经细胞主要包括神经元（neuron）、细胞核（nucleus）、轴突（axon）、树突（dendrites）以及突触。细胞核为神经细胞的主要处理机构；轴突为传递信号至其他神经元树突的主要介质；树突即树状传递线，专门接收来自其他神经元的信号；突触则是神经元间传递信号的连接点，如图 5.1 所示。

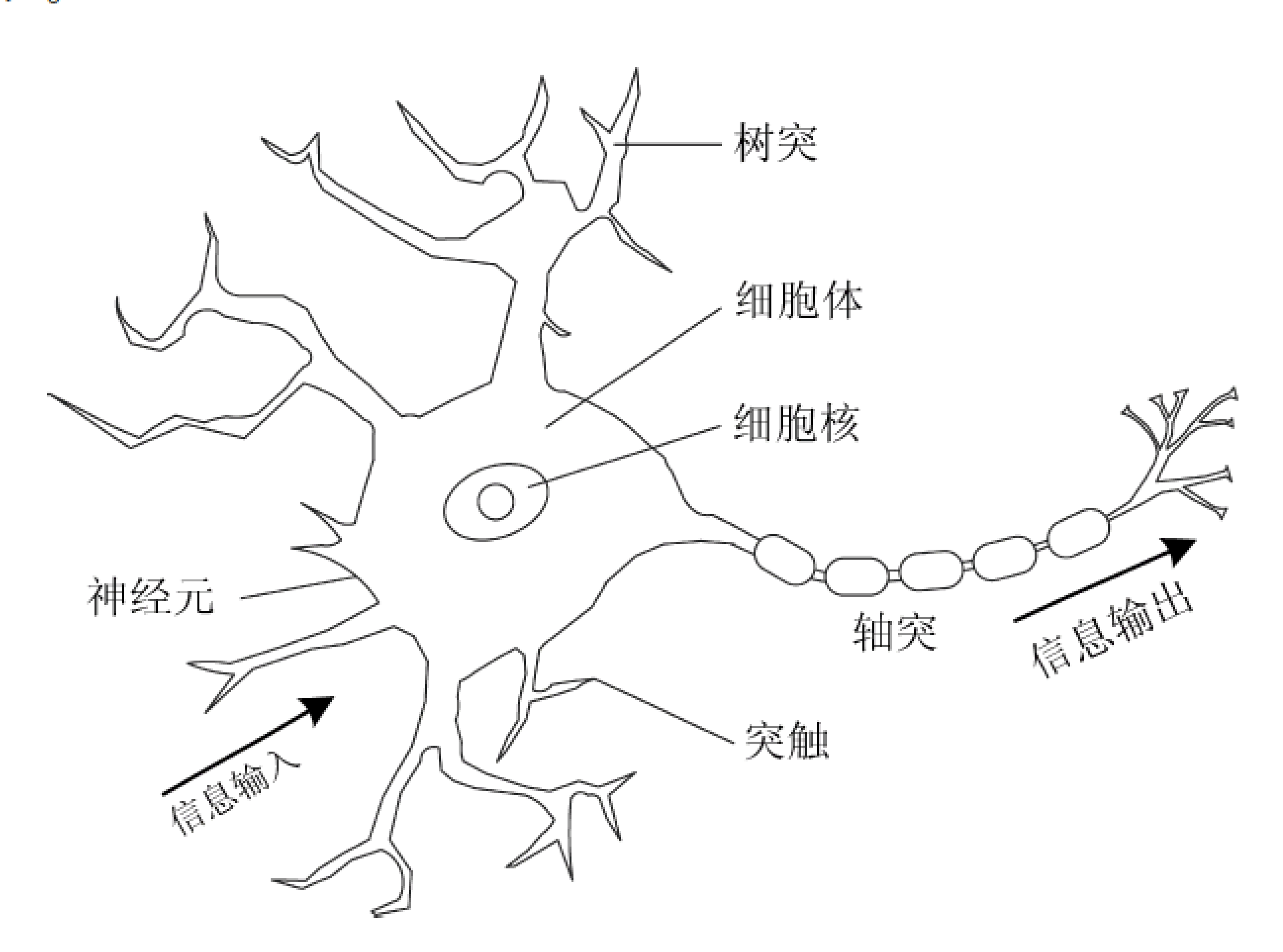


图 5.1　生物神经元架构

人工神经网络（artificial neural networks）模仿生物神经网络的信息处理系统，以处理复杂的问题，从其他人工神经元或外在环境取得信息，借由网络结构及不同的学习算法训练人工神经网络，使其输出能达到期望的目标。

人工神经网络分为不同阶段：①**学习（learning）**阶段主要是建立神经元间的连接模式、修正连接神经元之间的权重、调整神经元激活函数（activation function）中的门槛值；②**回想（recall）**阶段为当神经网络接收到一个输入的刺激后，依据建立的神经网络架构产生一个相应的输出值；③**归纳推演（induction）**阶段为从局部观察而推导出整体特性的过程，提供有效率的记忆与储存模式。

神经元是整个人工神经网络运作的基础，图 5.2 为一个神经元 k 的运算模型。假设有 p 个神经元输入信号 $x_i, i=1,2,\cdots,p$，至神经元 k，而第 $\bar{c}$ 个神经元对神经元 k 的连接关系与影响程度以权重 w_{ik} 表示，权重的大小表示神经元之间连接的强弱，若权重为正值，则表示该输入 x_i 为促进反应，反之，若权重为负号，则表示该输入 x_i 为抑制反应。对第 k 个神经元所接受的信号为所有输入信号 $\boldsymbol{X}=(x_0,x_1,x_2,\cdots,x_p)$ 与相对应权重 $\boldsymbol{W}=(w_{0k},w_{1k},w_{2k},\cdots,w_{pk})$ 的乘积加总 $net_k=\sum_{i=0}^{p}x_iw_{ik}$，而神经元的 w_{0k} 又称为门槛值(threshold)或偏误值(bias)。为了仿真神经元细胞接收信息后的作用，w_{0k} 的初始设定值会设定为负值，且 $x_0=1$，因此，根据所有输入信号的加权结果 $\sum_{i=1}^{p}x_iw_{ik}$ 与门槛值 x_0w_{0k} 相减的结果大于等于 0 或小于 0，会发出刺激或抑制的信号。最后神经元的输出会依据给定 net_k 下的函数值，如式(5.1)：

$$y_k = f(net_k) \tag{5.1}$$

函数 f 又称为激活函数，主要是用以转换 net_k 的函数。

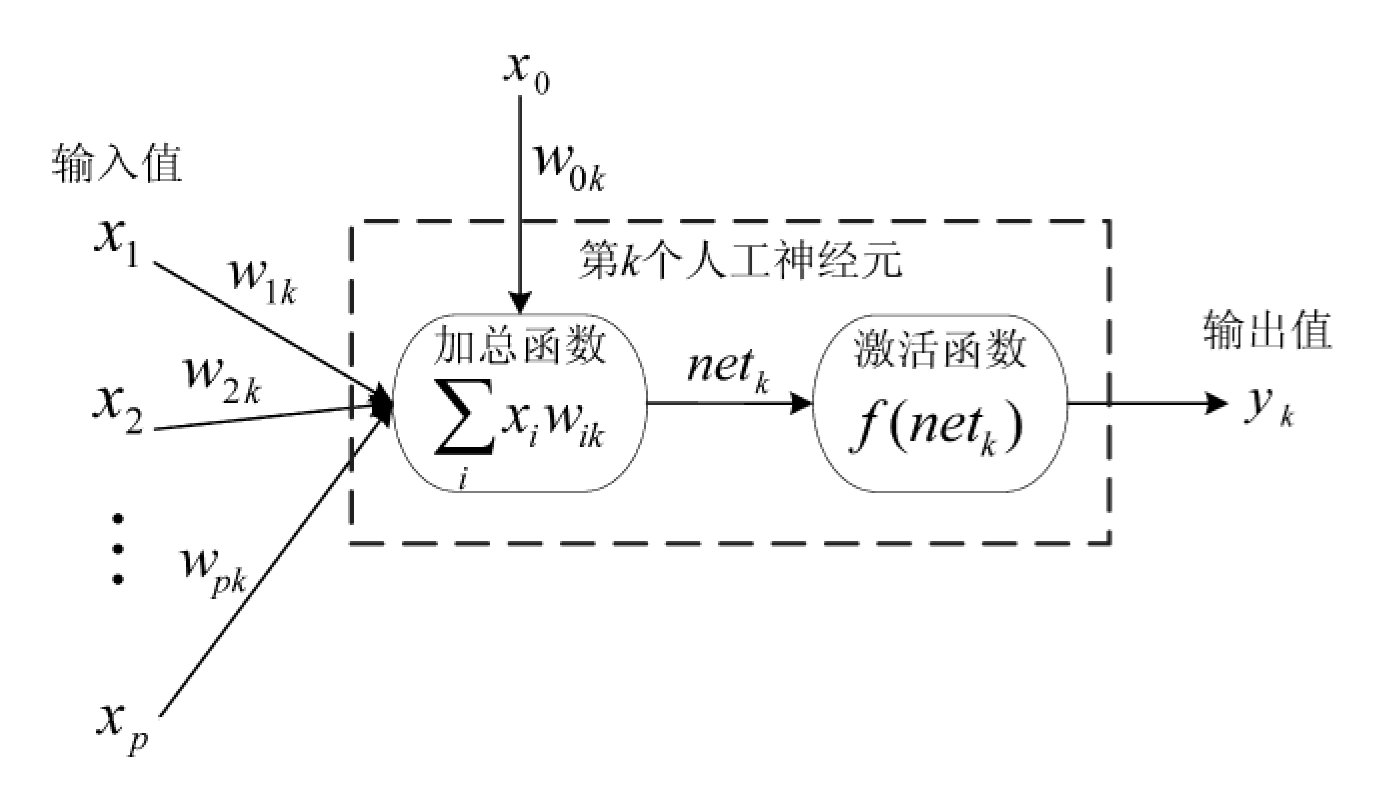


图 5.2 人工神经元运算模型

激活函数可借由线性或非线性转换 net_k 为神经元的输出值，表 5.1 为几种常见的激活函数。

表 5.1 激活函数

激活函数	数学式	函数图形
硬限幅函数 (hard limit function)	$y_k=\begin{cases}1, & net_k \geq 0\\ 0, & net_k < 0\end{cases}$	

续表

激活函数	数学式	函数图形
符号函数 (sign function)	$y_k=\begin{cases}1, & net_k\geqslant 0\\ -1, & net_k<0\end{cases}$	$y_k=f(net_k)$; net_k; -5 -4 -3 -2 -1 0 1 2 3 4 5; 1; -1
线性函数 (linear function)	$y_k=net_k$	$y_k=f(net_k)$; net_k; -5 -4 -3 -2 -1 0 1 2 3 4 5; 5 4 3 2 1 -1 -2 -3 -4 -5
S型函数 (sigmoid function)	$y_k=\dfrac{1}{1+e^{-net_k}}$	$y_k=f(net_k)$; net_k; 1; -5 -4 -3 -2 -1 0 1 2 3 4 5
双曲正切函数 (hyperbolic tangent function)	$y_k=\dfrac{e^{net_k}-e^{-net_k}}{e^{net_k}+e^{-net_k}}$	$y_k=f(net_k)$; net_k; 1; 0; -1; -5 -4 -3 -2 -1 0 1 2 3 4 5

假设一个神经元接收一输入向量 $\boldsymbol{X}=(1,3,6)$ 与其连接权重向量 $\boldsymbol{W}'=(-0.5,0.2,0.3)$，如图5.3所示，则该神经元的输出应该为所有输入值的加权结果 *net* 的函数，根据不同激活函数，其输出值的计算如下：

$net=1\times(-0.5)+3\times0.2+6\times0.3=1.9$；

硬限幅函数：$y=1$；

线性函数：$y=1.9$；

S 型函数：$y=\frac{1}{1+e^{-1.9}}=0.870$；

双曲线正切函数：$y=\frac{e^{1.9}-e^{-1.9}}{e^{1.9}+e^{-1.9}}=0.956$。

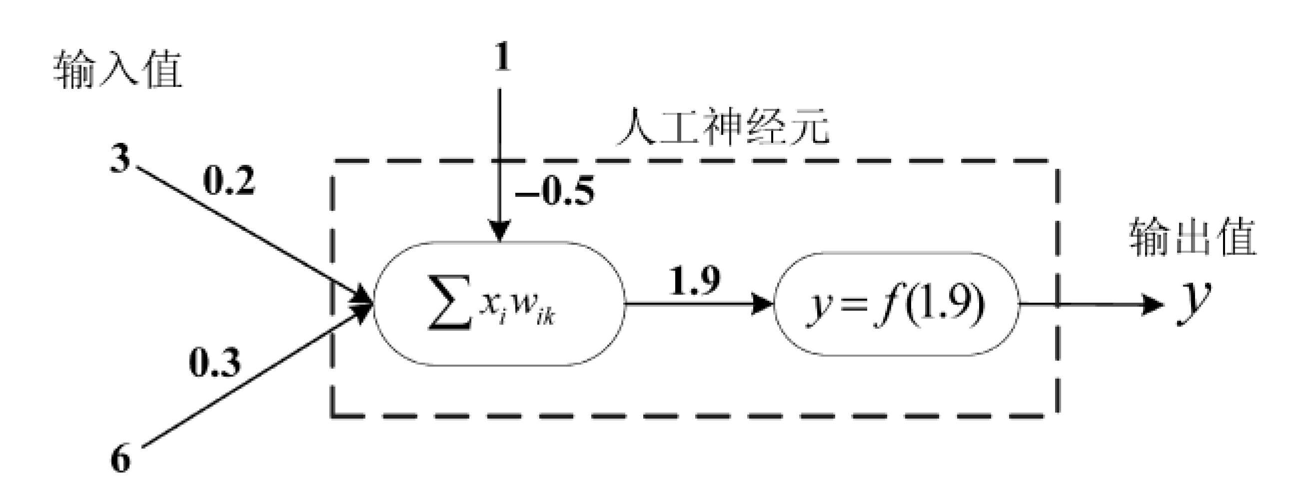


图 5.3　神经元计算范例

5.1　人工神经网络的基本结构

网络结构又称为网络拓扑(topology)，是由许多神经元或节点以各种连接方式所组成。图 5.4 为常见简单的三层网络拓扑，包括输入层、隐藏层与输出层，其定义与功用如下。

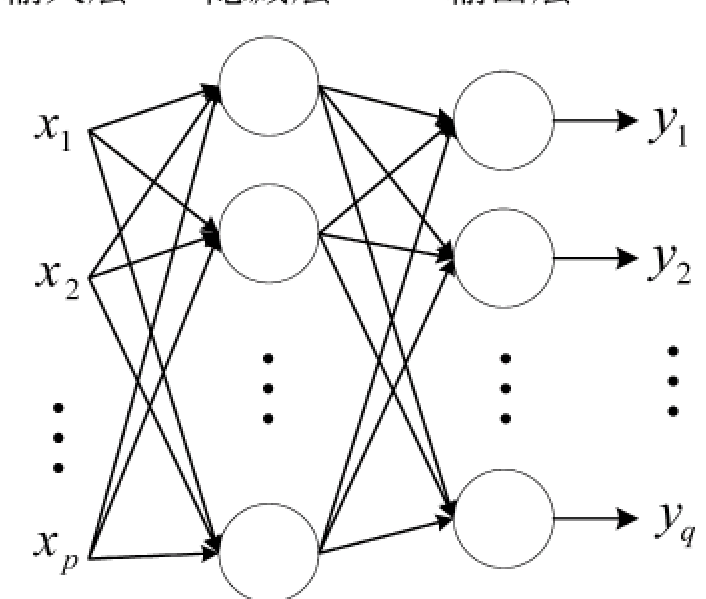


图 5.4　隐藏层的网络架构

1. 输入层(input layer)

输入层是处理单元接收外在环境所输入的信息，可依问题特性，常使用线性转换函数将输入数据转换成适应网络的信号。输入层的每个神经元只接收一个输入变量作为其输入值，并将输出值送至下一层中的各个神经元，因此输入层神经元的个数即等于输入变量的个数。输入层分成两种类型：①输入层中的神经元包含配重值、偏移量及转换函数；②输入层中的神经元则只具有接收输入变量的功能，输出值即等于输入变量，没有运算的功能。

2. 隐藏层(hidden layer)

隐藏层介于输入层与输出层之间，作为处理单元彼此间交互作用的内在结构解决非线性的问题。决定隐藏层神经元的个数并无特定规则，分析者可视数据复杂度调整隐藏层的层数(可以是零或是多层)与该隐藏层神经元的个数。通常须依赖经验、公式或以实验方式去决定其最适单元数目及使用的非线性转换函数。当隐藏层的数目为一层或两层时有较佳的收敛效果，若隐藏层神经元的数目过多，虽然能让训练集数据产生较小的误差值，但测试集数据的误差可能会不降反升，造成过度配适的现象。

3. 输出层(output layer)

输出层处理单元处理输出至外在环境的信息,处理单元的个数依不同问题而定,亦可使用非线性转换函数将输入数据转换成输出信号。输出层中每个神经元的输出值即网络的输出值,所以输出层神经元的个数等于网络的输出值个数。

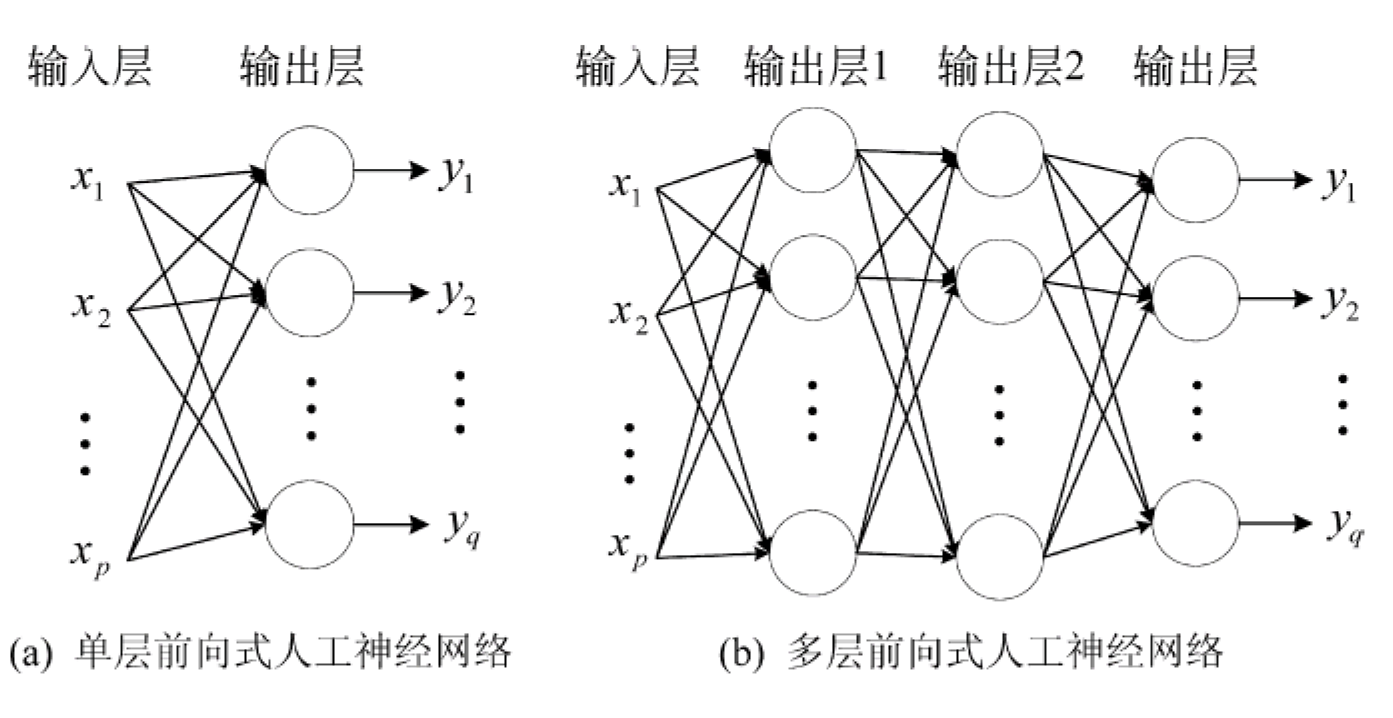
输出层的功能分为三种:①归一化输出:将同一层处理单元的原始输出值所组成的向量先行归一化,转化成单位长度向量后,再输出信号;②竞争化输出:从同一层处理单元的原始输出值组成的向量中,令一个或多个最强处理单元的输出值为 1(即优胜单元),其余处理单元的输出值为 0,再输出信号;③竞争化学习:从同一层处理单元的原始输出值组成的向量中,选择一个或多个最强势的处理单元,只调整与其相连的下层网络连接。

人工神经网络的学习能力与其系统架构的大小及形态有关,神经元个数太少将可能无法处理复杂的问题,神经元个数太多则可能导致过度配适。网络层数的选择并非越多越好,层数越多计算就越复杂,也就越容易出现局部优化的问题,因此,一般问题至多只要二层隐藏层即可。而在决定隐藏层内的神经元数时,可以采用尝试错误法(trial-and-error method)不断地递归测试,依不同问题的复杂度找出最佳处理单元数。

人工神经网络的依据连接架构可分为**前向式人工神经网络(feed-forward neural network)**与**反馈式人工神经网络(recurrent neural network)**。

前向式人工神经网络是由单层或多层的神经元所组成,神经元间的数据传递方向与整个网络的数据传递方向相同,为向前的单向传递,其信息传递的方式是从输入层(经由隐藏层)往输出层的方向传送,同侧间不相连且无递回传递,每一层神经元只会接收上层神经元所传送过来的输出值,并经过处理后得到一个新输出值。换言之,第一层隐藏层只会接收来自输入层的输入变量,而第二层隐藏层只会接收来自第一层隐藏层的输入变量。

就网络层而言,在处理简单的问题时并不需设定隐藏层,然而,当面临复杂且庞大的数据运算,或是单层网络架构所不能处理的异或函数(XOR)及复杂的非线性问题时,即须借助隐藏层的函数运算,处理交互作用,了解更多的高阶作用,以提升人工神经网络的效用。根据层数的多寡,又分为**单层前向式人工神经网络**与**多层前向式人工神经网络**,如图 5.5 所示。其中,单层前向式网络由于整个网络仅由一层具有信息处理能力的人工神经元所组成,功能性通常较差,只能处理线性问题;而多层前向式网络根据连接的连接方式又细分为"部



(a) 单层前向式人工神经网络　(b) 多层前向式人工神经网络

图 5.5　前向式人工神经网络

分连接"(partially connected)与"完全连接"(fully connected),其结构较为细密,因此可以处理较复杂的问题。

前向式人工神经网络常应用于图样识别(pattern recognition)、感知器(perceptron)、反向传播人工神经网络(back propagation neural network,BPNN)、线性联想记忆(linear associate memory)、自组织映射网络(self-organizing network)等。

反馈式网络架构至少有一反馈方向,可以递归给同一层或前一层的神经元,作为其输入数据,亦即此网络的输出会通过另一组连接值连接于网络的某处(如输入层或隐藏层),再反馈至网络本身,如图 5.6 所示。因此,反馈式人工神经网络为一种动态网络架构,在网络训练过程会通过神经元间的连接而显示不同的状态,直到达到平衡点,也就是当输入值改变或是已找出最佳参数组合与最佳模型时,才会停止动态移转。其常以多层网络架构呈现,以处理较复杂的问题。

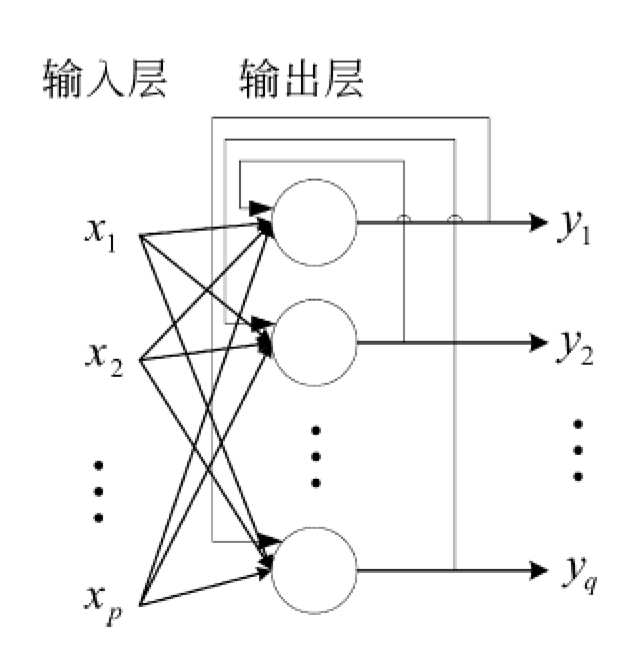


图 5.6 反馈式人工神经网络

反馈式网络主要是用来处理与时间有关的数据或问题,借由反馈的过程使得神经元间产生时间上的延迟,以加强网络的学习能力。应用反馈式人工神经网络架构的模型有自联想式记忆(auto-associative memory)、梯度搜寻法(gradient type)、暂时性关连式记忆(temporal associative memory)、自适应共振理论网络(adaptive resonance theory network,ART)等。

5.2 网络学习法

人工神经网络中的学习过程就是决定节点连接权重的过程。网络学习算法可以反复调整网络连接权重值,使神经网络的输出能达到最佳数值,神经元间的连接权重主要是经由训练组样本输入与输出值的结果逐步调整。网络学习方式可分为监督式学习(supervised learning)及无监督式学习(unsupervised learning)。

监督式学习在训练过程中会根据目标输出值调整权重大小,使得网络输出值与目标值的差异最小化;无监督式学习法则无目标可让网络产生的输出值对应比较,必须通过发掘与适应输入值所带来的信息,也就是从这些训练组样本中发掘出规则或是群类样型以建立模型。

通用学习算法(general learning rule)可用来说明权重调整的学习机制(Amari,1990)。假设有一输入样本 $\boldsymbol{X}=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$,权重 $\boldsymbol{W}_k=(w_{1k},w_{2k},\cdots,w_{pk})^{\mathrm{T}}$ 为所有连接到第 k 个节点的连接权重向量,而 w_{ik} 代表第 i 个输入单元连接至第 k 个节点的连接权重值,第 $t+1$ 次的权重值等于第 t 次权重值加上权重调整量,如式(5.2):

$$\boldsymbol{W}_k^{t+1}=\boldsymbol{W}_k^t+\Delta\boldsymbol{W}_k^t \tag{5.2}$$

其中,权重调整量 $\Delta\boldsymbol{W}_k^t$ 则是由输入样本向量 $\boldsymbol{X}$ 与学习信号(learning signal)e 决定,而学习信号又为权重向量、输入样本值、目标输出值($\boldsymbol{T}$)的函数 $e_k^t=f(\boldsymbol{W}_k^t,\boldsymbol{X}^t,\boldsymbol{T}_k^t)$,因此,权重 $\Delta\boldsymbol{W}_k^t$ 第 t 次的调整量 $\Delta\boldsymbol{W}_k^t$ 可定义如式(5.3):

$$\Delta \boldsymbol{W}_k^t = \eta e_k^t \boldsymbol{X}^t \tag{5.3}$$

其中，η 为学习率，学习率越大，则每次的权重调整量越大。

不同学习算法将分别应用在不同人工神经网络的权重调整与网络训练。如感知器学习法(perceptron learning method)、梯度下降学习法(gradient descent learning method)、随机性学习法(stochastic learning method)、竞争式学习法(competitive leaning method)、Hebbian 学习法(Hebbian learning method)等，可进一步参照(Hassoun，1995)与(Patternson，1996)。

感知器是人工神经网络学习算法中重要的基础算法，为监督式学习算法，主要用以解决分类问题。以**单层感知器(single layer perceptron)算法**为例，假设以符号函数作为激活函数，目标为降低目标值与神经元输出值的差异，$e_k = T_k - y_k$，$y_k = \text{sgn}(\boldsymbol{W}_k' \boldsymbol{X})$，所以各连接权重向量的调整量为 $\Delta \boldsymbol{W}_k = \eta[T_k - \text{sgn}(\boldsymbol{W}_k' \boldsymbol{X})]\boldsymbol{X}$，所以根据此调整量规则，当目标值与神经元输出值有差异时即进行修正。

[范例 5.1]，如表 5.2 所示，说明如何利用单层感知器学习算法更新权重，若 $\boldsymbol{W}^1 = [-0.8 \quad 0.5 \quad 0.5]$，学习率 $\eta = 0.1$，$x_0 = 1$，其连接权重更新的过程如下。

表 5.2　[范例 5.1]的数据

Input		Output
x_1	x_2	T
1	1	1
−1	1	1
1	−1	1
−1	−1	−1

当 Input$\boldsymbol{X} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$，$net = \boldsymbol{W}'\boldsymbol{X} = 1 \times (-0.8) + 1 \times 0.5 + 1 \times 0.5 = 0.2$，因为 $y_1 = \text{sgn}(0.2) = 1$，$T_1 = 1$，$T_1 - y_1 = 0$，所以不需更新连接权重，$\boldsymbol{W}^2 = \boldsymbol{W}^1$。

当 Input$\boldsymbol{X} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$，$net = \boldsymbol{W}'\boldsymbol{X} = 1 \times (-0.8) + (-1) \times 0.5 + 1 \times 0.5 = -0.8$，因为 $y_2 = \text{sgn}(-0.8) = -1$，$T_2 = 1$，$T_2 - y_2 = 2$，所以连接权重需进行更新为 $\boldsymbol{W}^3$：

$$\boldsymbol{W}^3 = \boldsymbol{W}^2 + \Delta \boldsymbol{W}^2 = \begin{bmatrix} -0.8 \\ 0.5 \\ 0.5 \end{bmatrix} + 0.1 \times 2 \times \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.6 \\ 0.3 \\ 0.7 \end{bmatrix}$$

当 Input$\boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$，$net = \boldsymbol{W}'\boldsymbol{X} = 1 \times (-0.6) + (-1) \times 0.3 + 1 \times 0.7 = -0.2$，因为 $y_3 = \text{sgn}(-0.2) = -1$，$T_3 = 1$，$T_3 - y_3 = 2$，所以连接权重需进行更新为 $\boldsymbol{W}^4$：

$$W^4 = W^3 + \Delta W^3 = \begin{bmatrix} -0.6 \\ 0.3 \\ 0.7 \end{bmatrix} + 0.1 \times 2 \times \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -0.4 \\ 0.5 \\ 0.5 \end{bmatrix}$$

当 Input$\boldsymbol{X}=\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$，$net=\boldsymbol{W}'\boldsymbol{X}=1\times(-0.4)+(-1)\times0.5+(-1)\times0.5=-1.4$，因为 $y_4=\text{sgn}(-1.4)=-1$，$T_4=-1$，$T_4-y_4=0$，所以连接权重无须更新。

在模型训练阶段，连接权重 $\boldsymbol{W}$ 会不断调整，直到神经元输出值与目标值一致为止。在单层感知器的学习过程中，如果欲解决的问题为线性可分割，也就是仅用线性函数即可分割两类，如[范例 5.1]的问题，但如果$(x_1,x_2,T)=(1,1,1)$变成$(x_1,x_2,T)=(1,1,-1)$，则该问题就变成一个 XOR 的分类问题，需用非线性函数才能正确的分割为两类。解决的办法就是由单层人工神经网络改为多层人工神经网络，多层人工神经网络的学习算法可利用反向传播学习算法(back-propagation learning method)的算法，详细说明请见 5.3 节。

5.3 反向传播人工神经网络

反向传播人工神经网络是广为使用的监督式学习网络(Rumelhart & McClelland，1986)。一般使用反向传播学习算法与多层感知器架构即称为**反向传播人工神经网络(back-propagation neural network，BPNN)**。

反向传播人工神经网络所使用的符号如下：

- i　　输入层的第 i 个节点，$i=1,2,\cdots,p$
- j　　隐藏层的第 j 个节点，$j=1,2,\cdots,h$
- k　　输出层的第 k 个节点，$k=1,2,\cdots,q$
- l　　第 l 个训练数据，$l=1,2,\cdots,n$
- w_{ji}　　连接输入层的节点 i 与隐藏层的节点 j 的权重值
- w_{kj}　　连接隐藏层的节点 j 与输出层的节点 k 的权重值
- x_i^l　　第 l 笔训练数据的节点 i 的输入值
- z_j^l　　第 l 笔训练数据在隐藏层节点 j 的神经元输出值
- y_k^l　　第 l 笔训练数据在输出层节点 k 的神经元输出值
- d_k^l　　第 l 笔训练数据在输出层节点 k 的目标值
- f　　神经元的激活函数
- w_{j0}　　隐藏层节点 j 连接的门槛值
- w_{k0}　　输入层节点 k 连接的门槛值
- η　　学习率
- δ_k　　输出层节点 k 的误差量
- m　　学习循环

5.3.1 网络架构

反向传播人工神经网络通常采用前向式多层网络模式，其基本架构如图 5.7 所示，包括

输入层、隐藏层与输出层。

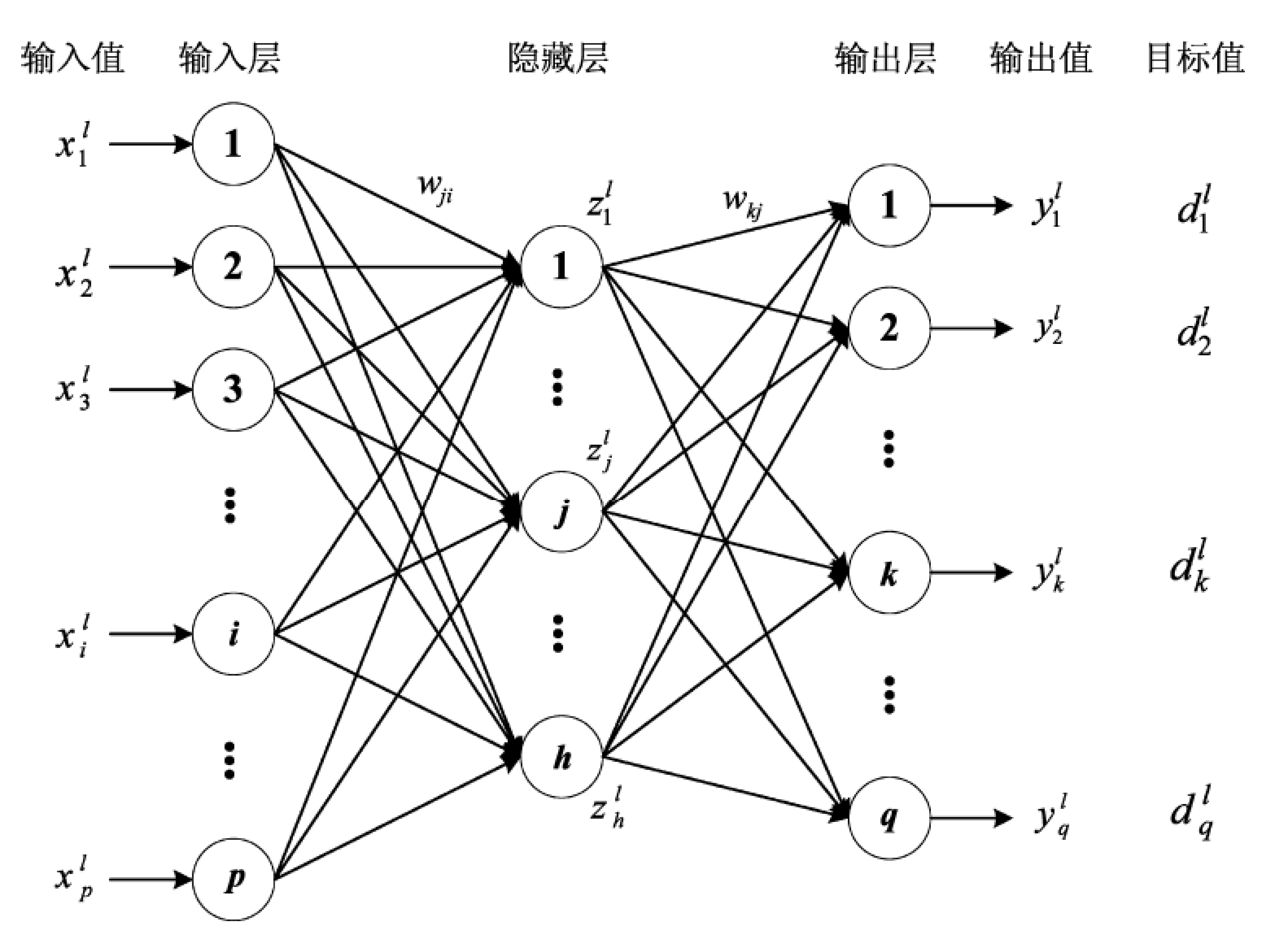


图 5.7　反向传播人工神经网络架构

(1) **输入层**：即网络的输入变量，神经元数目的多寡可视输入属性个数而定。为提升网络的训练效率，往往会事先进行训练组样本数据的前置处理，以达到更好的效率。

(2) **隐藏层**：用于处理输入层单元间的交互作用或非线性的关系。隐藏层数目往往借由尝试错误法所决定，并以一层或两层时收敛效果为佳，以避免过度配适。

(3) **输出层**：用来代表人工神经网络的输出值，神经元数目的多寡需视问题而定，而输出神经元可依问题需求而决定是否要采用介于 1 和 －1 间的线性或非线性双弯曲转换函数。

反向传播神经网络的学习算法使用的是误差反向传播算法，演算过程包括正向及反向的传递。在正向向前传递的过程，是将输入信号经由网络内部的权重及门槛值处理后，再传递至隐藏层，在隐藏层将所有传来的信息通过转换函数转换成一输出值，最后再传向输出层。因此，代表输入层的神经元会直接影响到隐藏层的神经元，进而间接影响到输出层的神经元，因而得到输入与输出间的相互关系。如果在神经元的输出值与目标值不一致，则反向向后传递，将计算值与目标值的误差信号沿着原来网络连接的通路返回，根据学习法则沿途修正网络内各层的权重及门槛值，更新后的内部各权重值将作为新的连接权重，再输入下一笔数据重新进行正向及反向的运算，所以称做“反向传播学习算法”。如此经过的经由多笔数据迭代后，直到神经元输出值将趋近于目标值或达到最大周期数，当输入所有训练数据集的数据进入网络并完成学习的过程即称为一个周期。

完整网络学习需要不断地重复学习，也就是说如果训练数据有 100 个样本数据，最大学习周期数为 100，则最大的网络学习则将输入 10 000 笔。若训练结果不理想，则可尝试增加训练周期，同时依照问题的复杂程度不同，不断尝试以找出每次学习循环是否均须依相同的次序输入训练范例，亦或随机挑选。

5.3.2 学习算法

反向传播网络的学习算法包括正向向前传递与反向向后传递两种过程，向前传递中，输入信息从输入层通过隐藏层加权计算，经激活函数转换处理后，最后传向输出层并计算出网络输出值，当网络输出值与目标值有所差异时，则向后传递误差信息，修改各层神经元的权重与各神经元的门槛值，以修正输出层神经元输出值与目标值的差距。反向传播人工神经网络的学习是基于最陡下降法(gradient steepest descent method)通过迭代使训练数据目标值与网络输出值误差最小化的过程。

假设输入层与隐藏层、隐藏层与输出层间均为完全连接，以一层隐藏层为例，说明反向传播算法权重的更新。在隐藏层与输出层的输出值是经由激活函数所计算而得，如式(5.4)与式(5.5)：

$$z_j = f(net_j) = f\left(\sum_i w_{ji} x_i\right) \tag{5.4}$$

$$y_k = f(net_k) = f\left(\sum_j w_{jk} z_j\right) \tag{5.5}$$

在网络训练过程中，是以目标值与网络输出值误差极小化为目标来调整网络各节点连接的权重值。定义每一笔数据下，其误差值 E 为所有输出层节点的输出值与目标值的误差平方和，如式(5.6)所示。若网络的输出值与实际目标值的差异越小，则表示网络学习的效果越好。

$$E = \frac{1}{2}\sum_{k=1}^{q} (d_k - y_k)^2 \tag{5.6}$$

反向传播学习算法的目的就是调整权重使得误差值 E 最小，误差值的大小主要受到输出层的输出值 y_k 的影响，也就是各连接权重的影响，因此可借由调整权重值以最小化误差平方合，可利用坡度下降学习法调整权重连接值的大小，其调整的幅度取决于学习率 η 的设定大小如式(5.7)所示：

$$\Delta W = -\eta \frac{\partial E}{\partial W} \tag{5.7}$$

隐藏层与输出层间的连接权重调整，以及输入层与隐藏层间的连接权重调整说明如下。

1. 隐藏层与输出层的连接权重调整

隐藏层与输出层的连接权重的调整可根据误差函数 E 对 w_{kj} 的偏微分 $\partial E/\partial w_{kj}$ 用微积分的连锁律求得，如式(5.8)：

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial net_k}\frac{\partial net_k}{\partial w_{kj}} \tag{5.8}$$

其中，

$$\frac{\partial net_k}{\partial w_{kj}} = \frac{\partial}{w_{kj}}\sum_j w_{kj} z_j = z_j \tag{5.9}$$

$$\frac{\partial E}{\partial net_k} = \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial net_k} = -(t_k - y_k) f'(net_k) \tag{5.10}$$

再假设变量 δ_k 为输出层第 k 个输出神经元的误差量，如式(5.11)：

$$\delta_k = (t_k - y_k) f'(net_k) = (t_k - y_k) y_k (1 - y_k) \tag{5.11}$$

因此综合式(5.8)至式(5.11)，隐藏层第 j 个节点与输出层第 k 个节点的连接权重可改为式(5.12)：

$$\Delta w_{kj} = \eta \delta_k z_j \tag{5.12}$$

2. 输入层与隐藏层的连接权重调整

输入层与隐藏层的连接权重调整可根据误差函数 E 对 w_{ji} 的偏微分 $\partial E/\partial w_{ji}$ 用微积分的连锁律求得，如式(5.13)：

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ji}} \tag{5.13}$$

其中，

$$\frac{\partial net_j}{\partial w_{ji}} = \frac{\partial}{w_{ji}}\sum_i w_{ji}x_i = x_i \tag{5.14}$$

$$\frac{\partial z_j}{\partial net_j} = f'(net_j) \tag{5.15}$$

而误差函数对隐藏层节点 j 的偏微分可利用连锁律求解，如式(5.16)：

$$\frac{\partial E}{\partial z_j} = \sum_k\left(\frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial net_k}\frac{\partial net_k}{\partial z_j}\right) = \sum_k -(t_k - y_k)f'(net_k)w_{jk} \tag{5.16}$$

再假设变量 δ_j 为隐藏层第 j 个输出神经元的误差量，如式(5.17)：

$$\delta_j = \sum_k(\delta_k w_{kj})f'(net_j) = \sum_k(\delta_k w_{kj})z_j(1-z_j) \tag{5.17}$$

综合式(5.13)至式(5.17)，输入层第 i 个节点与隐藏层第 j 个节点的连接权重可改为式(5.18)：

$$\Delta w_{ji} = -\eta\frac{\partial E}{\partial w_{ji}} = \eta\delta_j x_i \tag{5.18}$$

5.3.3 反向传播人工神经网络步骤

反向传播人工神经网络的训练过程可分为 10 个步骤：

(1) 设定网络结构、输入层、隐藏层、输出层节点个数，以及学习率、最大学习周期等参数，设定 $l=1$。

(2) 随机乱数生成初始权重 w_{ji} 与 w_{kj}，选定节点输出转换的激活函数。

(3) 随机选取一训练样本组，包括输入数据向量 $\boldsymbol{x}_i^l=(x_1^l, x_2^l, \cdots, x_p^l)$ 与目标值向量 $\boldsymbol{d}_i^l=(d_1^l, d_2^l, \cdots, d_q^l)$。

(4) 计算隐藏层每个节点的输出值 z_j^l，以及输出层每个节点的输出值 y_k^l。

(5) 计算误差函数 E。

(6) 计算输出层的差距量 δ_k 与隐藏层的差距量 δ_j。

$$\delta_k = (d_k - y_k)y_k(1-y_k)$$

$$\delta_j = \sum_k(\delta_k w_{kj})z_j(1-z_j)$$

(7) 计算输出层与隐藏层间的连接权重修正量 Δw_{kj}^l，以及隐藏层与输入层间的连接权重修正量 Δw_{ji}^l。

(8) 更新连接权重。

$$w_{kj}^{l+1} = w_{kj}^{l} + \Delta w_{kj}^{l}$$
$$w_{ji}^{l+1} = w_{ji}^{l} + \Delta w_{ji}^{l}$$

(9) $l=l+1$，重新回到步骤(3)，直到所有训练组数据均输入完成。

(10) 重新回到步骤(2)到步骤(9)，直到达到设定的最大周期数。

5.3.4　反向传播人工神经网络范例

根据三层的网络模式架构，如图 5.8 所示，欲将两组训练数据(x_1，x_2)与其目标值(d_1，d_2)，依照反向传播算法进行模式训练，其演算过程如下。

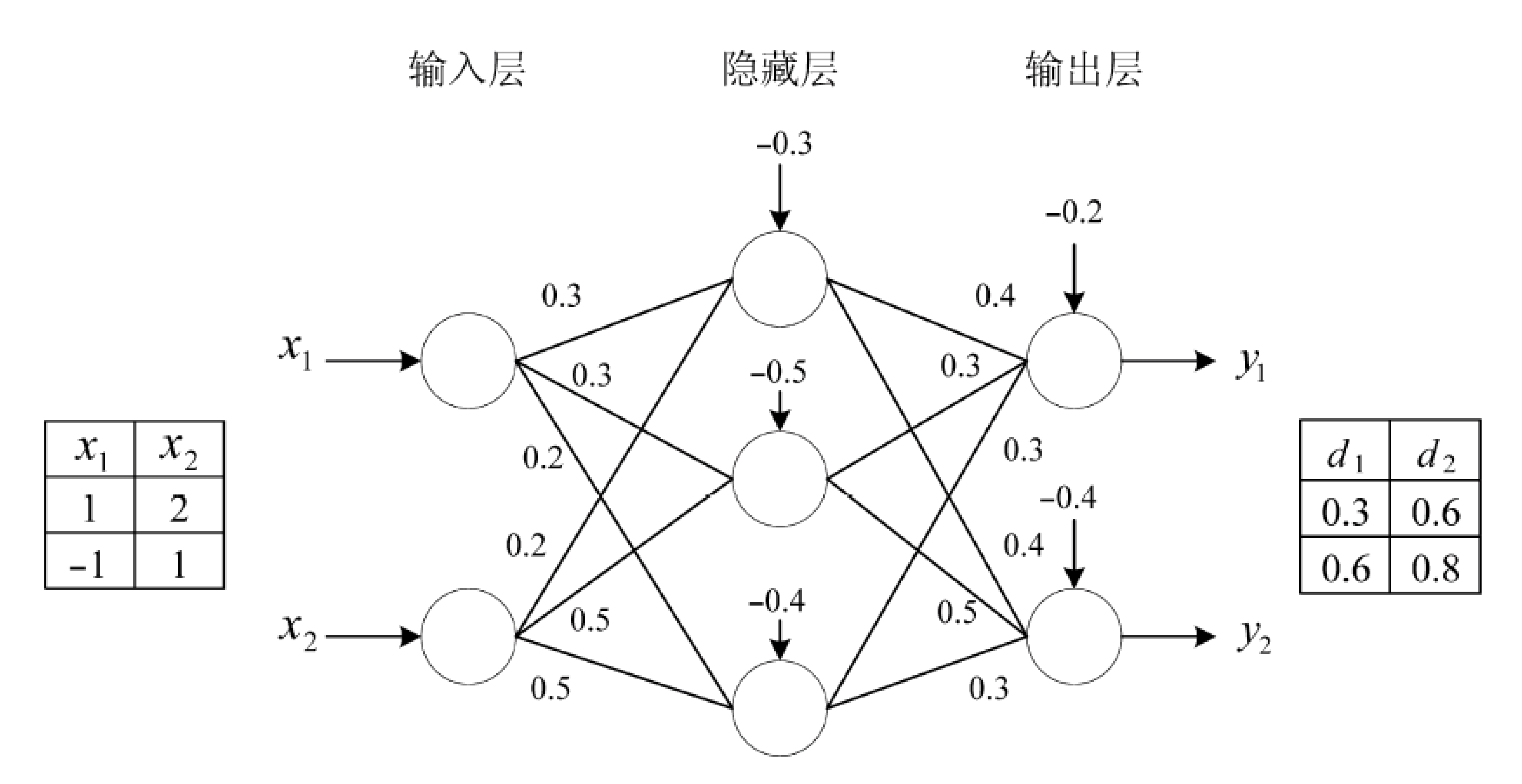


图 5.8　反向传播人工神经网络范例

(1) 设定学习率 η 为 0.2，输入层、隐藏层、输出层的节点数分别为 2、3、2。

(2) 以随机乱数初始网络误差值 w_{ji}、w_{kj} 及初始权重 w_{j0} 与 w_{k0}，见图 5.8 所示，所用的激活函数为 $f=\dfrac{1}{1+\mathrm{e}^{-net}}$，$l=1$。

(3) 输入第一笔训练数据(x_1，x_2)=(1，2)，(d_1，d_2)=(0.3，0.6)。

(4) 计算隐藏层每个节点的输出值(z_1^1，z_2^1，z_3^1)=(0.60，0.69，0.69)，以及输出层每个节点的输出值(y_1^1，y_2^1)=(0.66，0.69)。

(5) 计算误差函数 $E=\dfrac{1}{2}[(0.3-0.66)^2+(0.6-0.69)^2]=0.0689$。

(6) 计算输出层的差距量 δ_k 与隐藏层的差距量 δ_j。

$$\delta_{k=1} = (0.3-0.66)\times 0.66(1-0.66) = -0.0808$$
$$\delta_{k=2} = (0.6-0.69)\times 0.69(1-0.69) = -0.0193$$
$$\delta_{j=1} = (-0.0808\times 0.4+(-0.0193)\times 0.4)\times 0.6\times(1-0.6) = -0.01$$
$$\delta_{j=2} = (-0.0808\times 0.3+(-0.0193)\times 0.5)\times 0.69\times(1-0.69) = -0.007$$
$$\delta_{j=3} = (-0.0808\times 0.3+(-0.0193)\times 0.3)\times 0.69\times(1-0.69) = -0.006$$

(7) 计算输出层与隐藏层间的连接权重修正量 Δw_{kj}^{l}。

$$\Delta w_{10} = 0.2\times(-0.0808) = -0.016,\quad \Delta w_{20} = 0.2\times(-0.0193) = -0.004$$
$$\Delta w_{11} = 0.2\times(-0.0808)\times 0.6 = -0.010$$
$$\Delta w_{21} = 0.2\times(-0.0193)\times 0.6 = -0.002$$

$$\Delta w_{12} = 0.2 \times (-0.0808) \times 0.69 = -0.011$$
$$\Delta w_{22} = 0.2 \times (-0.0193) \times 0.69 = -0.003$$
$$\Delta w_{13} = 0.2 \times (-0.0808) \times 0.69 = -0.011$$
$$\Delta w_{23} = 0.2 \times (-0.0193) \times 0.69 = -0.003$$

计算隐藏层与输入层间的连接权重修正量 Δw_{ji}^{l}。

$\Delta w_{10} = 0.2 \times (-0.01) = -0.002$，　$\Delta w_{20} = 0.2 \times (-0.007) = -0.0014$

$\Delta w_{11} = 0.2 \times (-0.01) \times 1 = -0.002$，　$\Delta w_{21} = 0.2 \times (-0.007) \times 1 = -0.0014$

$\Delta w_{12} = 0.2 \times (-0.01) \times 2 = -0.004$，　$\Delta w_{22} = 0.2 \times (-0.007) \times 2 = -0.0028$

$\Delta w_{30} = 0.2 \times (-0.006) = -0.0012$

$\Delta w_{31} = 0.2 \times (-0.006) \times 1 = -0.0012$

$\Delta w_{32} = 0.2 \times (-0.006) \times 2 = -0.0024$

(8) 更新输出层与隐藏层间的连接权重。

$w_{10}^{2} = (-0.3) + (-0.016) = -0.316$，　$w_{20}^{2} = (-0.4) + (-0.004) = -0.404$

$w_{11}^{2} = (0.4) + (-0.01) = 0.39$，　$w_{21}^{2} = (0.4) + (-0.002) = 0.389$

$w_{12}^{2} = (0.3) + (-0.011) = 0.289$，　$w_{22}^{2} = (0.5) + (-0.003) = 0.497$

$w_{13}^{2} = (0.3) + (-0.011) = 0.289$，　$w_{23}^{2} = (0.3) + (-0.003) = 0.297$

更新隐藏层与输入层间的连接权重。

$w_{10}^{2} = (-0.3) + (-0.002) = -0.302$，　$w_{20}^{2} = (-0.5) + (-0.0014) = -0.5014$

$w_{11}^{2} = (0.3) + (-0.002) = 0.298$，　$w_{21}^{2} = (0.3) + (-0.0014) = 0.2986$

$w_{12}^{2} = (0.2) + (-0.004) = 0.196$，　$w_{22}^{2} = (0.5) + (-0.0028) = 0.4972$

$w_{30}^{2} = (-0.4) + (-0.0012) = -0.4012$

$w_{31}^{2} = (0.2) + (-0.0012) = 0.1988$

$w_{32}^{2} = (0.5) + (-0.0024) = 0.4976$

(9) $l=l+1$，重新回到步骤(3)，再输入第 2 笔训练数据$(x_1,x_2)=(-1,1)$，$(d_1,d_2)=(0.6,0.8)$，直到所有训练组数据均输入完成。

(10) 重新回到步骤(2)到步骤(9)，直到达到设定的最大周期数。

5.4　自组织映射网络

自组织映射图(self-organizing map，SOM)网络属于非监督式的学习算法，又称 Kohonen 网络(Kohonen，1995)，采用竞争式的网络架构，其输出层的神经元会根据输入数据特征，在输出空间中呈现出有意义的拓扑结构，亦即使任意维度的输入向量映射至二维或低维度的映射网络图上，也就是将输入数据空间以非线性的投影法转换至二维特征的空间上。由于所产生的拓扑结构可反映输入数据本身的特征，因而称做“自组织映射网络图”(Kraaijveld *et al*.，1995)。

SOM 能借由网络架构发掘数据本身的特征与关联性，聚集特征相近的数据，进而分群。SOM 的特性是物以类聚，能够处理大量且高维度的多变量数据，且能保留原始数据所隐含的重要信息。SOM 在网络学习的过程中为采用竞争式学习算法，首先将输出神经元安排在

有前后关系的直线或平面上(基本上为二维平面),通过输入向量量化与投影,将多维度的数据映射到输出层的拓扑坐标上,以视觉上容易检查的二维网络拓扑方式呈现其群聚结果。基本的运作原理为计算出各输入特征值映射至输出层的每一神经元之距离,如欧式距离(Euclidean distance),再比较所有的距离以选出最小距离值的神经元为优胜神经元。根据竞争式学习算法,胜出的网络输出神经元连接权重会越来越强,并调整获胜输出神经元周围相邻近的神经元的连接权重,使其更接近原始的输入向量,以减少与输入向量间的距离,逐渐形成各群聚区域。

5.4.1 网络架构

SOM 网络所使用的符号以及表示法的定义如下:

i 输入层的第 i 个节点,$i=1,2,\cdots,p$

k 输出层的第 k 个节点,$k=1,2,\cdots,q$

l 第 l 个训练数据,$l=1,2,\cdots,n$

t 网络学习迭代次数,最大迭代次数为输入训练样本数

$\boldsymbol{X}$ SOM 网络的输入向量,$\boldsymbol{X}=(x_1,x_2,\cdots,x_p)$

x_i^l 第 l 组的训练样本组,常以向量方式表示输入样本,其中,其中,$i=1,2,\cdots,p$

w_{ki} 第 k 个神经元与第 i 个输入神经元的连接权重值

D_k^l 第 l 笔数据的输入向量$\boldsymbol{X}^l$ 与第 k 个神经元的连接权重值向量$\boldsymbol{W}_k$ 的欧式距离 $\|\boldsymbol{X}^l-\boldsymbol{W}_k\|$

T 规定的最大迭代次数

m 第 m 次学习循环

R_t 第 t 次迭代时的邻近半径值

η_t 第 t 次迭代时的学习率值

自组织映射网络的网络架构如图 5.9 所示,有别于反向传播人工神经网络架构,SOM 网络架构仅包含输入层与输出层。

(1) **输入层**:输入层主要是借由加载输入变量为输入神经元,来自输入向量 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_p\}$,其中每一个神经元皆相互独立,且连接权重也相互独立。输入层的神经元数目的多寡主要依据问题而定。

(2) **输出层**:输出层主要为神经元的输出,输出层神经元并非只有一个,其数目的多寡需视问题而定,并会以一维向量或二维拓扑图呈现,图 5.9 为一个二维网络拓扑图。每一个输出单元都会连接到所有输入单元,并以连接权重作为神经元之间关系的强弱。不同于其他人工神经网络,SOM 在输出层加入了网络拓扑及邻近区域的概念。

(3) **网络拓扑**:网络拓扑的组成不限任意形状及任意维度,其输出层的处理单元排列方式可为一维或多维空间的形态,且形状可为矩形、三角形、圆形,甚至是任意形状等。SOM 输出神经元间的相对位置具有意义,并根据具有相同特征的输入向量,而用二维的拓扑结构形态显示数据间的群聚关系,如图 5.10 所示。

(4) **拓扑坐标**:拓扑坐标是用以决定输出单元在网络拓扑的相对位置,在 SOM 中用以计算各输出神经元的邻近关系。举例而言,二维的网络拓扑均会有对应的二维拓扑坐标。拓扑坐标会因定位点的取法不同而将神经元标识成不同的几何坐标,如图 5.10,以最左下

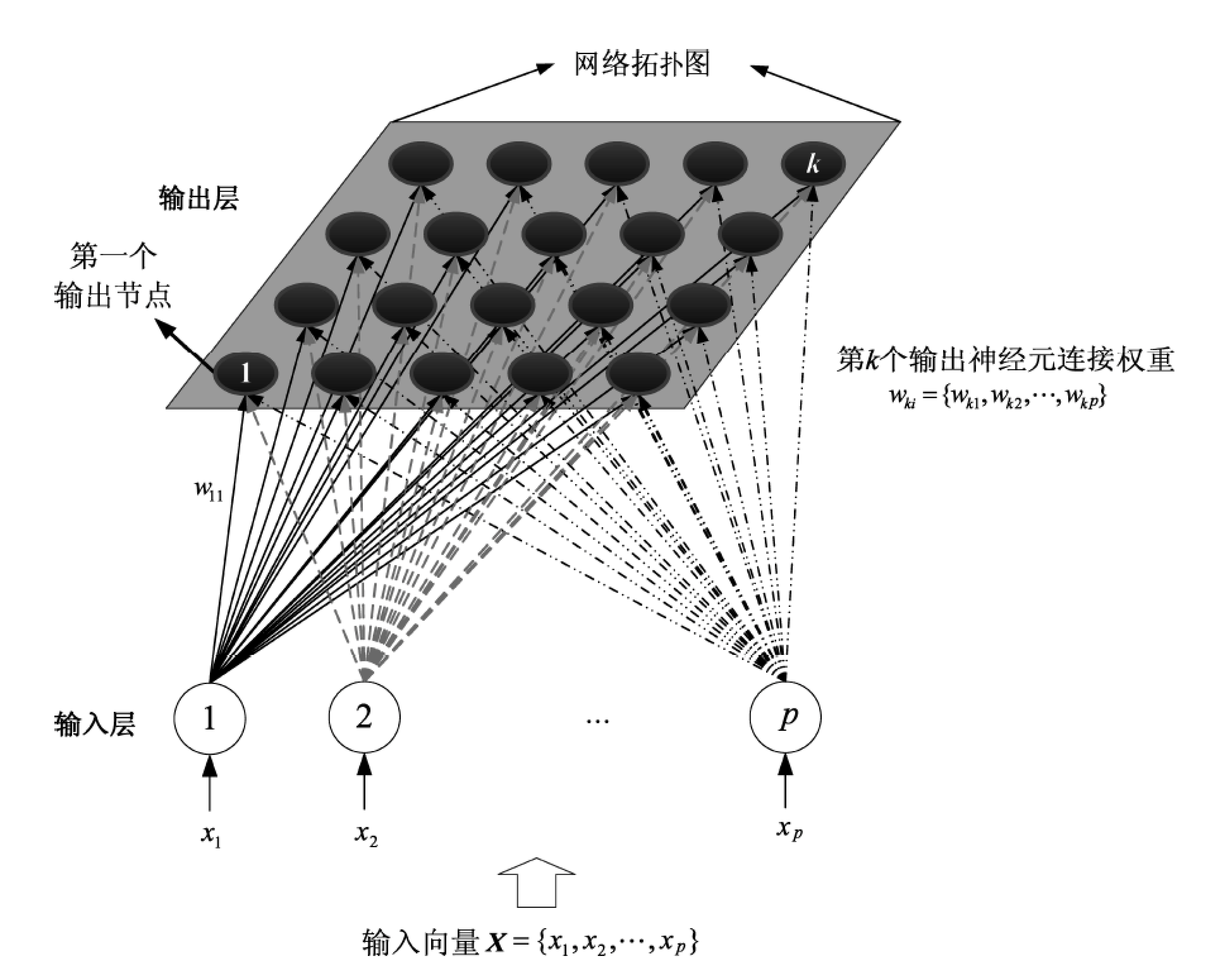


图 5.9　自组织映射网络的网络基本架构(数据源：修改自 Patterson，1996)

角的神经元为坐标原点(0,0)，则点Ⅰ的坐标将会表示为(1,2)且点Ⅱ的坐标会表示为(4,1)。

为了得到有意义的二维拓扑映射图，SOM 网络在学习过程中，除了调整获胜神经元的连接权重外，其周围附近的神经元也会一并被调整。如同手指被针刺到后，感到疼痛的不仅是刺到的点，连同附近皮肉组织也会有疼痛的感觉，而距离越远的部位则越没有感觉。

计算神经元的邻近关系主要根据邻近函数的结果，与邻近函数相关的参数包括邻近中心、邻近区域、邻近半径、邻近距离，其关系如图 5.11 所示。

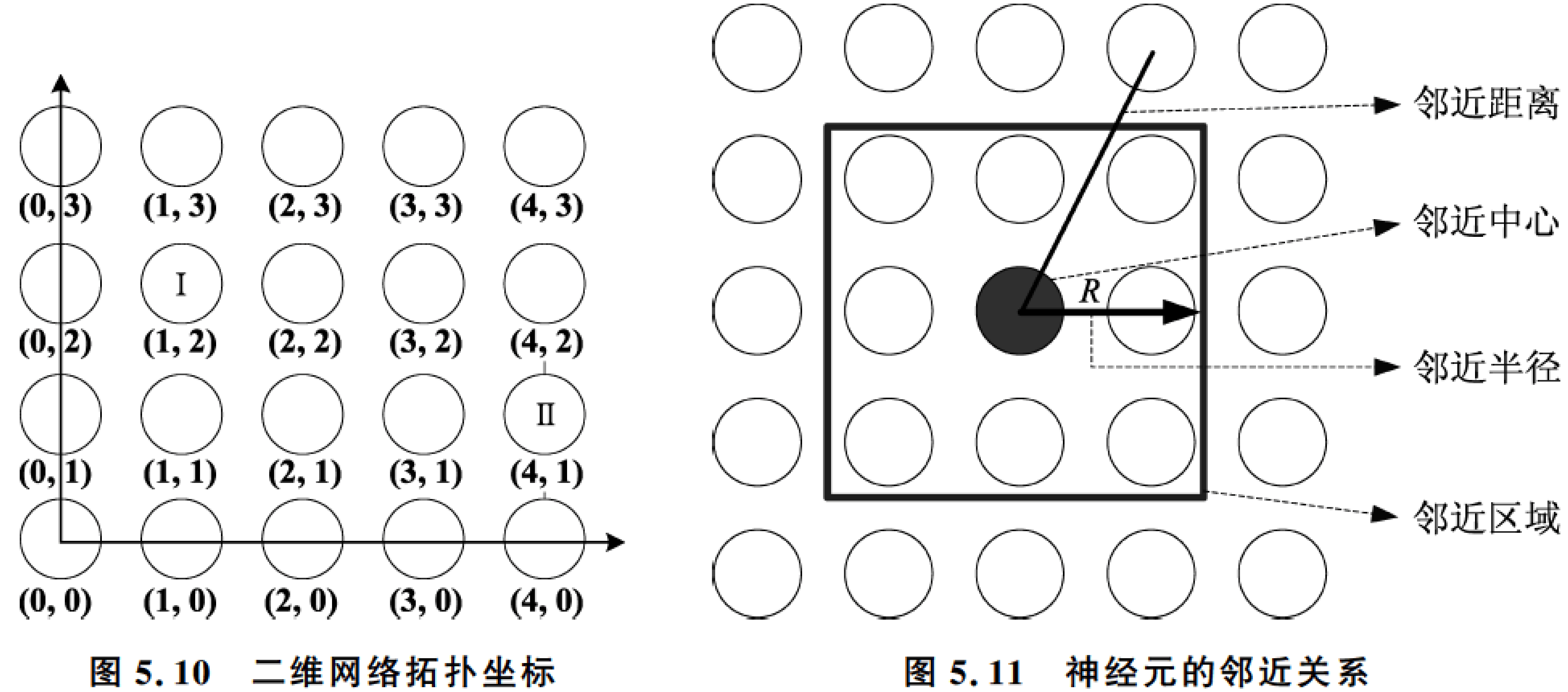


图 5.10　二维网络拓扑坐标

图 5.11　神经元的邻近关系

(5) **邻近中心**：为控制邻近函数中心位置的参数，一般以网络拓扑中胜出的神经元为邻近中心，如图 5.11 中所述，邻近中心的决定可利用式(5.19)：

$$D_k = \min_k \| \boldsymbol{X} - \boldsymbol{W}_k \| \tag{5.19}$$

(6) **邻近半径**：控制邻近区域大小的参数，以 R 表示，初始邻近半径的设定会比较大，再借由学习循环次数逐渐缩小，如图 5.12 所示，若 R^m 代表第 m 次学习循环(epoch)时的邻近半径，则第 $m+1$ 次的邻近半径为 $R^{m+1}=\lambda R^m$，其中，λ 为调整系数，$0<\lambda<1$。

(7) **邻近区域**：网络拓扑中，以邻近中心为主，在邻近半径长度范围内的区域，邻近区域可用不同形状，常见的如矩形、六角形等，随着网络学习的过程中会逐渐缩小，如图 5.12 所示。

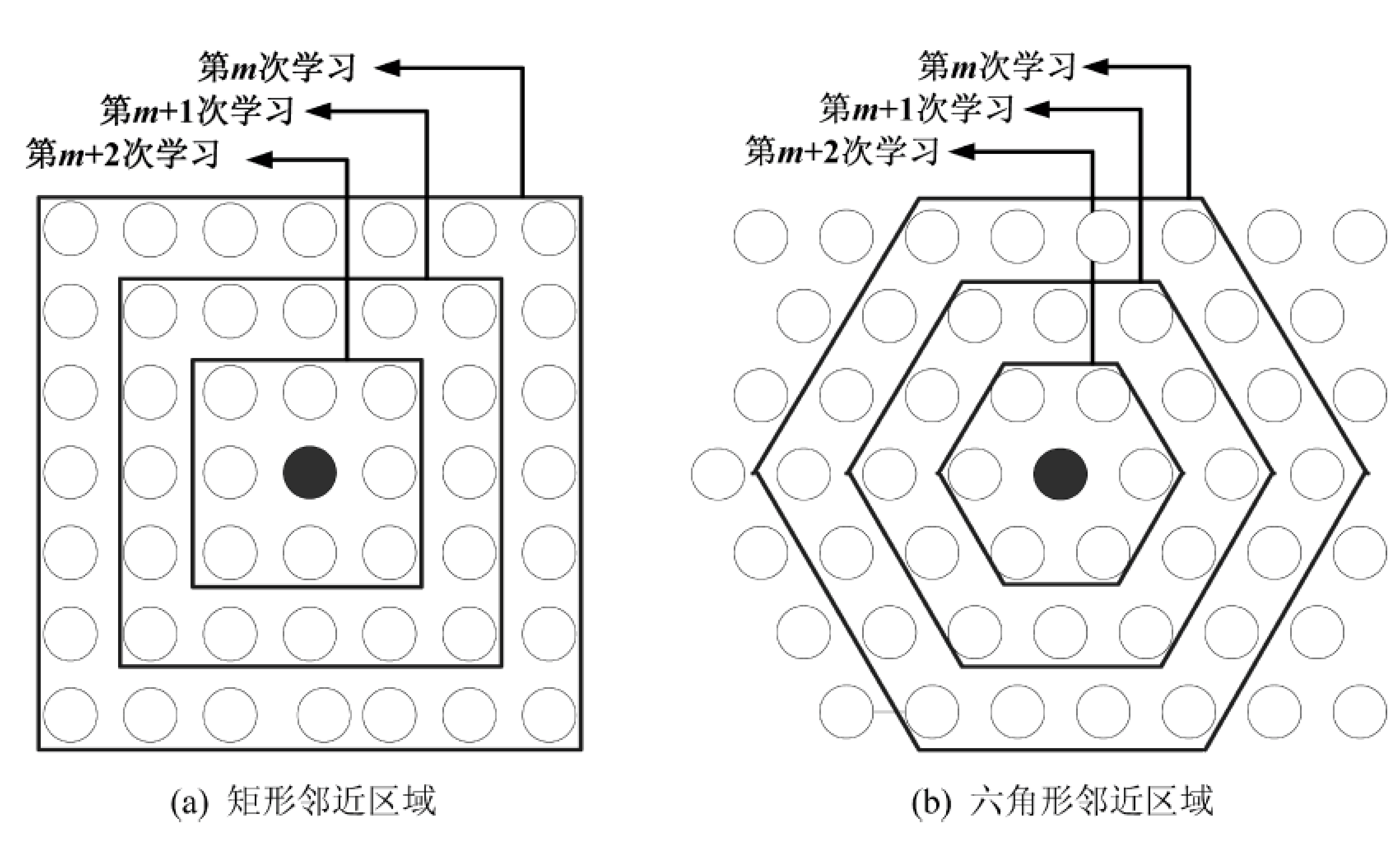


图 5.12 邻近神经元与邻近半径示意图

(8) **邻近距离**：代表输出神经元 k 在网络拓扑中与邻近中心 v 的距离，邻近距离的计算是根据拓扑坐标定义，而以欧式距离计算其距离，如式(5.20)：

$$D_{vk} = \| u_k - u_v \| \tag{5.20}$$

其中，$\| \cdot \|$ 为 norm，u 代表神经元在拓扑坐标的位置，u_v 表示邻近中心的坐标值，u_k 表示拓扑结构上神经元的坐标值。

(9) **邻近函数**：表示神经元 k 与邻近中心 v 的邻近关系值，如式(5.21)：

$$\delta_{vk} = \mathrm{e}^{-(D_{vk}/R)^2} \tag{5.21}$$

邻近神经元 k 的权重更新结果可依据其调整量 Δw_k，如式(5.22)与式(5.23)：

$$\Delta w_k = \eta \delta_{vk} \| \boldsymbol{X} - \boldsymbol{W}_v \| \tag{5.22}$$

$$w_k^{t+1} = w_k^t + \Delta w_k \tag{5.23}$$

5.4.2 学习算法

SOM 主要利用迭代的方式计算各输入向量与输出层处理单元间的连接权值向量，通过竞争式学习算法，不断调整连接权重值使其越接近原输入向量的值，直到输入向量与连接权重的总距离为最小时或最大学习循环时，方停止训练。

SOM 网络的学习过程中，每一笔输入训练数据会通过连接权重的大小，找到与该输入

向量最近似的神经元作为优胜神经元，因此，可定义误差函数为输入向量与连接权重向量的距离。

$$E = \min_k \| \boldsymbol{X} - \boldsymbol{W}_k \| = \min_k \left[\frac{1}{2} \sum_{i=1}^{p} (x_i - w_{ki})^2 \right] \tag{5.24}$$

同样可利用最陡下降法求得使得误差函数最小的调整权重连接值的大小，其调整的幅度取决于学习率 η 的设定大小如式(5.25)所示：

$$\Delta w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}} \tag{5.25}$$

将式(5.24)代入式(5.25)可得

$$\Delta w_{ki} = -\eta \frac{\partial}{\partial w_{ki}} \left(\min_k \frac{1}{2} \sum_{i=1}^{p} (x_i - w_{ki})^2 \right) \tag{5.26}$$

若神经元 k 刚好与输入向量间为最小距离，则进行权重调整，其调整量为

$$\Delta w_{ki} = -\eta \frac{\partial}{\partial w_{ki}} \left(\min_{\forall k} \frac{1}{2} \sum_{i=1}^{p} (x_i - w_{ki})^2 \right) = -\eta \frac{1}{2} [-2(x_i - w_{ki})] = \eta (x_i - w_{ki}) \tag{5.27}$$

若神经元 k 与输入向量间并非最小距离，则其之间的连接权重不进行调整，$\Delta w_{ki}=0$。此外，根据定义的邻近关系，优胜神经元周围的神经元与输入向量的连接权重也会一并更新，如式(5.22)所示。若邻近距离越大，连接加权值修正也越小。

在 SOM 算法中，决定停止网络训练的标准有许多种，例如达到最大学习循环次数，或是输入向量与连接权重的总距离小于门槛值。然而，大部分的应用时仍多以达到最大循环次数作为最后学习停止的条件。

5.4.3　SOM 人工神经网络步骤

SOM 网络的训练过程分为 7 个步骤：

(1) 随机产生与设定初始连接权重值向量 $\boldsymbol{W}_k = \{w_{k1}, w_{k2}, \cdots, w_{kp}\}$，设定网络拓扑大小与输出层节点个数，设定 $l=1$。

(2) 决定邻近半径的初始值 R，设定学习率的初始设定值 η、最大学习循环次数。

(3) 随机选取一训练样本组 $\boldsymbol{X}_i^l = (x_1^l, x_2^l, \cdots, x_p^l)$，根据式(5.24)求得输入的训练样本的优胜神经元。

(4) 更新与此优胜神经元相连接的权重值，以及与此优胜神经元邻近区域神经元所连接的权重值、权重值的更新方式。

$$w_k^{t+1} = w_k^t + \Delta w_k$$
$$\Delta w_k = \eta \delta_{vk} \| \boldsymbol{X} - \boldsymbol{W}_v \|$$

(5) $l=l+1$，回到步骤(3)，直到所有训练组数据均输入完成。

(6) 调整邻近半径 R 与学习率 η。

(7) 重新回到步骤(3)到步骤(5)，直到达到设定的最大周期数。

5.4.4　自组织映射图网络范例

假设欲将四个输入向量(1,1,0,0)、(0,0,1,1)、(0,0,0,1)、(1,0,0,0)，利用 SOM 算法

进行分群，分群个数为两群，则演算过程如下所示。

(1) 随机给予连接权重值向量，并进行初始化为

$$W=\begin{bmatrix}0.2 & 0.7\\0.7 & 0.5\\0.5 & 0.3\\0.6 & 0.8\end{bmatrix}$$

因为输入层为四个神经元所组成，输出层至多分成两群，故仅由两个神经元组成，设定 $l=1$。

(2) 设定邻近半径 $R=0$，采用赢者全拿的竞争学习机制，并不考虑对邻近神经元的权重值进行更新。设定学习率的初始设定值 $\eta=0.7$。

(3) 输入第 1 笔数据，第一组训练数据 $\boldsymbol{X}^1=(1,1,0,0)^{\mathrm{T}}$。计算连接输入层的各节点至输出层节点 1 与节点 2 的欧式距离：

$$D_1^1=\|\boldsymbol{X}^1-\boldsymbol{W}_1\|=\sqrt{\sum_{i=1}^{4}(x_i-w_{1i})^2}$$
$$=\sqrt{(1-0.2)^2+(1-0.7)^2+(0-0.5)^2+(0-0.6)^2}=1.158$$
$$D_2^1=\|\boldsymbol{X}^1-\boldsymbol{W}_2\|=\sqrt{\sum_{i=1}^{4}(x_i-w_{2i})^2}$$
$$=\sqrt{(1-0.7)^2+(1-0.5)^2+(0-0.3)^2+(0-0.8)^2}=1.034$$

因此可找出输出层的节点 2 具有最小距离，故为优胜神经元。

(4) 更新与此优胜神经元所连接的权重值如下：

$$\boldsymbol{W}_2^2=\boldsymbol{W}_2^1+0.7\times(\boldsymbol{X}^1-\boldsymbol{W}_2^1)$$
$$=\begin{bmatrix}0.7\\0.5\\0.3\\0.8\end{bmatrix}+0.7\times\begin{bmatrix}0.3\\0.5\\-0.3\\-0.8\end{bmatrix}=\begin{bmatrix}0.91\\0.85\\0.09\\0.24\end{bmatrix}$$

因此可得更新后的连接权重矩阵为$\boldsymbol{W}^2=\begin{bmatrix}0.20 & 0.91\\0.70 & 0.85\\0.50 & 0.09\\0.60 & 0.24\end{bmatrix}$。

(5) $l=l+1$，回到步骤(3)，

输入第 2 笔训练数据，$\boldsymbol{X}^2=(0,0,1,1)^{\mathrm{T}}$。

(3) 计算连接输入层的各节点至输出层节点 1 与节点 2 的欧式距离：

$$D_1^2=\sqrt{(0-0.20)^2+(0-0.70)^2+(1-0.50)^2+(1-0.60)^2}=0.97$$
$$D_2^2=\sqrt{(0-0.91)^2+(0-0.85)^2+(1-0.09)^2+(1-0.24)^2}=1.719$$

输出层的节点 1 具有最小距离，故为优胜神经元。

(4) 更新与此优胜神经元所连接的权重值，如下：

$$\boldsymbol{W}_1^3=\boldsymbol{W}_1^2+0.7\times(\boldsymbol{X}^2-\boldsymbol{W}_1^2)$$

$$=\begin{bmatrix}0.20\\0.70\\0.50\\0.60\end{bmatrix}+0.7\times\begin{bmatrix}-0.20\\-0.70\\0.50\\0.40\end{bmatrix}=\begin{bmatrix}0.06\\0.21\\0.85\\0.88\end{bmatrix}$$

可得更新后的连接权重矩阵为$\boldsymbol{W}^3=\begin{bmatrix}0.06&0.91\\0.21&0.85\\0.85&0.09\\0.88&0.24\end{bmatrix}$。

(5) $l=3$,回到步骤(3),

输入第 3 笔训练数据,$\boldsymbol{X}^3=(0,0,0,1)^{\mathrm{T}}$。

(3) 计算连接输入层的各节点至输出层节点 1 与节点 2 的欧式距离:

$$D_1^3=\sqrt{(0-0.06)^2+(0-0.21)^2+(0-0.85)^2+(1-0.88)^2}=0.886$$

$$D_2^3=\sqrt{(0-0.91)^2+(0-0.85)^2+(0-0.09)^2+(1-0.66)^2}=1.294$$

输出层的节点 1 具有最小距离,故为优胜神经元。

(4) 更新与此优胜神经元所连接的权重值,如下:

$$\boldsymbol{W}_1^4=\boldsymbol{W}_1^3+0.7\times(\boldsymbol{X}^3-\boldsymbol{W}_1^3)$$

$$=\begin{bmatrix}0.06\\0.21\\0.85\\0.88\end{bmatrix}+0.7\times\begin{bmatrix}-0.06\\-0.21\\-0.85\\0.12\end{bmatrix}=\begin{bmatrix}0.018\\0.063\\0.255\\0.964\end{bmatrix}$$

可得更新后的连接权重矩阵为$\boldsymbol{W}^4=\begin{bmatrix}0.018&0.91\\0.063&0.85\\0.255&0.09\\0.964&0.24\end{bmatrix}$。

(5) $l=4$,回到步骤(3),

输入第 4 笔训练数据,$\boldsymbol{X}^4=(1,0,0,0)^{\mathrm{T}}$。

(3) 计算连接输入层的各节点至输出层节点 1 与节点 2 的欧式距离:

$$D_1^4=\sqrt{(1-0.018)^2+(0-0.063)^2+(0-0.255)^2+(0-0.964)^2}=1.401$$

$$D_2^4=\sqrt{(1-0.91)^2+(0-0.85)^2+(0-0.09)^2+(0-0.24)^2}=0.892$$

输出层的节点 2 具有最小距离,故为优胜神经元。

(4) 更新与此优胜神经元所连接的权重值,如下:

$$\boldsymbol{W}_2^5=\boldsymbol{W}_2^4+0.7\times(\boldsymbol{X}^4-\boldsymbol{W}_2^4)$$

$$=\begin{bmatrix}0.91\\0.85\\0.09\\0.24\end{bmatrix}+0.7\times\begin{bmatrix}0.09\\-0.85\\-0.09\\-0.24\end{bmatrix}=\begin{bmatrix}0.973\\0.255\\0.027\\0.072\end{bmatrix}$$

可得更新后的连接权重矩阵为$W^5=\begin{bmatrix}0.018 & 0.973\\ 0.063 & 0.255\\ 0.255 & 0.027\\ 0.964 & 0.072\end{bmatrix}$。

(5) 所有训练数据均输入至 SOM 网络学习。

(6) 调整邻近半径与学习率。

(7) 重新输入训练数据,并回到步骤(3)至步骤(5),直到最大学习循环。

5.5 自适应共振理论人工神经网络

自适应共振理论(adaptive resonance theory,ART)的神经网络模型(Grossberg,1987;Grossberg,1976)为无监督式学习算法,主要应用于辨识及分群,可利用输入图样与储存记忆的相似度(matching score)来完成此一任务。ART 网络是一种动态架构的神经网络,克服了一般竞争学习网络在输入图样重复的情况下所产生的不稳定现象(即输入相同图样时,可能于此一迭代会分在 C 类,而于下一次迭代则分到 D 类)。ART 人工神经网络是一个实时系统,能够对任意序列的输入图样,组织成稳定的辨识码(recognition code),因此其演算过程具有适应性机制,可避免产生网络不稳健的状态。

所谓的共振理论是指,在竞争学习下优胜的神经元必须符合原输入样型,才有资格进行更新以及学习,并通过网络的前向与反馈的路径所产生的交互作用,来监视系统的学习行为,更新权值向量,使得输出图样能重复出现,以达到共振状态(resonant state)。由此发展出来的算法有 ART1、ART2、ART3 和 Fuzzy ART 等多种模式。ART1 只适用于输入值为二元变量值(Carpenter & Grossberg,1987a);ART2 则可用在输入值为连续性数值(Carpenter & Grossberg,1987b);ART3 为阶层式的 ART 网络架构,以化学发散的概念使搜寻过程更有效率(Carpenter & Grossberg,1990);Fuzzy ART 则是合并 ART 算法与模糊算法机制(Carpenter *et al.*,1991)的模式。表 5.3 列出此四种相关模型的特性并加以比较。

表 5.3 ART 相关模型理论

模 型	特 性
ART1 (Grossberg,1976)	用在二元值(0 与 1)的图样识别上
ART2 (Carpenter & Grossberg,1987b)	针对模拟图样识别
ART3 (Carpenter & Grossberg,1990)	具有平行搜寻能力的阶层式架构
Fuzzy ART (Carpenter *et al.*,1991)	保有输入图样信号大小的信息,为一种将 ART1 模型与模糊集合理论结合的网络,允许输入向量扩展至[0,1]之间的模糊数(fuzzy number)

人类的记忆系统具有保留与储存已知事物的功能,当记忆新事物时,与原有记忆可能产生矛盾,因此需要良好的记忆系统来区隔及学习旧有记忆以及吸收新事物。此系统必须符合两个条件:稳定性(stability)及可塑性(plasticity)。一个实时学习系统需要有足够的稳定性来抗拒环境中不相干的事物或干扰以适当地保留旧事物,但又要有足够的可塑性来因应环境快速地改变与学习新事物。然而,因新旧事物的门槛值规定不易,此两种特性有时相

辅相成，有时又互相矛盾。

自适应共振理论采用人类记忆系统的运作方式，以**警戒值测试（vigilance test）**权衡稳定性与可塑性，以建立良好记忆系统与评估新旧事物的机制。警戒值门槛值的设定会影响到输入图样的辨识结果，当门槛值越大，输入图样与旧有记忆的储存图样间的匹配（match）程度就越高，所得的分类结果也越相似；反之，当门槛值设越小，匹配程度就越低，分类结果就越不一致。换言之，警戒值门槛值设的高低将控制网络的“稳定性”与“可塑性”，警戒值越高，网络的可塑性越高；反之，警戒值设定的越低，网络也就越具稳定性。

5.5.1　网络架构

ART 网络使用的符号及表示法如下：

i　　输入层的第 i 个节点，$i=1,2,\cdots,p$

k　　输出层的第 k 个节点，$k=1,2,\cdots,q$

l　　输入层的训练样本组数，$l=1,2,\cdots,n$

$\boldsymbol{I}$　　二元值输入的特征向量 $\boldsymbol{I}=[I_1,I_2,\cdots,I_p]$，其中，$I_i\in\{0,1\}$，$i=1,2,\cdots,p$

$\boldsymbol{X}$　　特征检测区 F_1 的状态（activation）向量 $\boldsymbol{X}=[x_1,x_2,\cdots,x_p]$

$\boldsymbol{Y}$　　接收区 F_2 的状态向量 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_q]$

$\boldsymbol{S}$　　特征检测区 F_1 的输出信号向量 $\boldsymbol{S}=[s_1,s_2,\cdots,s_p]$

$\boldsymbol{U}$　　接收区 F_2 的输出信号向量 $\boldsymbol{U}=[u_1,u_2,\cdots,u_q]$

$\boldsymbol{V}$　　由接收区 F_2 至特征检测区 F_1 的输入信号 $\boldsymbol{V}=[v_1,v_2,\cdots,v_p]$

w_{ki}^b　　由特征检测区 F_1 往接收区 F_2 的权重值向量

w_{ik}^t　　由接收区 F_2 往特征检测区 F_1 的权重值向量

ρ　　警戒参数值，$0<\rho<1$

自适应共振理论的网络架构如图 5.13 所示，如同 SOM 网络架构，仅包含输入层与输出层。

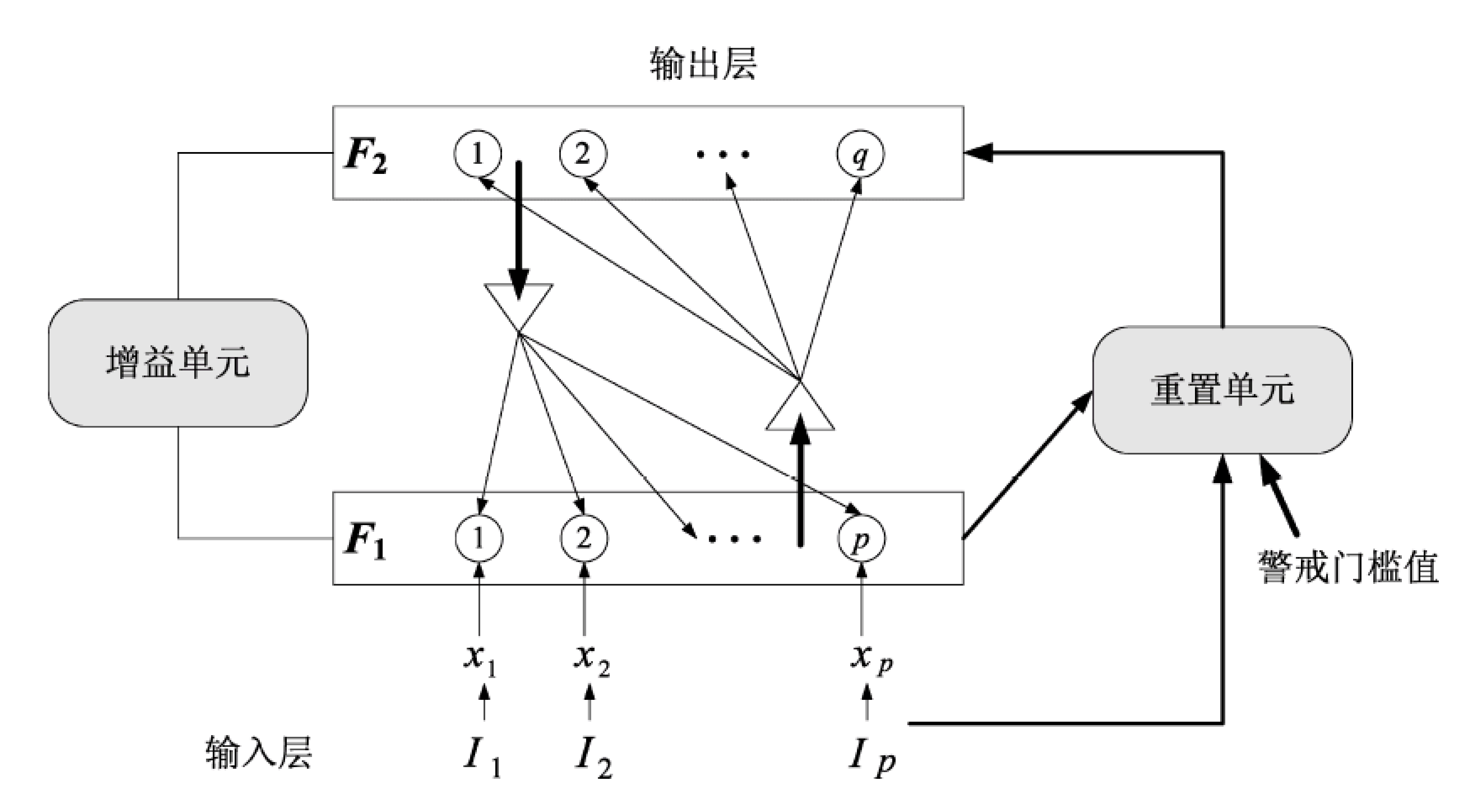


图 5.13　自适应共振理论网络架构（数据源：修改自 Patterson，1996）

ART 网络的演算机制为将 p 维度的输入向量映射至单一的输出分类。ART 的网络架构分为两个层次：输入层与输出层。这两个层次间的节点完全连接，包含前向连接与反馈

连接(Barto *et al.*,1983)。此三种组件的组成可以快速搜寻、比对与匹配出近似输入图样的种类,在不扰乱旧有记忆下,学习并记忆新图样。

(1) **输入层**:即为样本数据的输入向量,其处理单元个数与数据特征数有关。ART1 人工神经网络为 ART 网络模型的最早变形,其输入向量仅限于二元变量值;ART2 的输入向量则为连续性数值。

(2) **输出层**:用以表现网络的输出变量,每一个输出神经元即代表一个分群图样,和自组织映射网络的输出层定义类似,差别在于后者有“网络拓扑”与“邻近区域”的观念,但自适应共振理论网络则无。其输出层的处理单元数目最初只有一个,在学习过程中会逐渐增加,最后稳定在一定的数目,学习过程即告结束,此和其他人工神经网络模式输出层单元的数目为固定值极为不同。

(3) **网络连接**:自适应共振理论的每一个输入层单元与输出层单元间有前向与反馈两方向的网络连接,由下往上的连接是负责让输入层通过权值 b_{ij} 的计算,并输送至输出层竞争;由上往下的连接则是负责让优胜神经元的图样形态输送回输入层比对,若比对的结果相似则更新旧有记忆,否则须另建立新的图样群组,以储存新的记忆。

输入向量自输入层进入网络架构中,需经过两项测试:

(1) **相似度测试**:此为由下往上的搜寻比对,通过权重 w_{ki}^b 运算公式与输出层的旧有神经元记忆组进行比对,相似度最高的神经元 k 即为优胜神经元 k^*,如式(5.28)与式(5.29):

$$net_k = \sum_{i=1}^{p} w_{ki}^b x_i \tag{5.28}$$

$$net_{k^*} = \max_k (net_k) \tag{5.29}$$

(2) **警戒值测试**:此为由上往下的搜寻比对,有时相似度最大者不一定能通过警戒值测试,因此,为确保所建立的网络模式的效率,会进行再次检验,通过权重 $w_{ik^*}^t$ 用以计算该优胜神经元 k^* 需同时具有最大的相似度,如式(5.30):

$$\frac{\|\boldsymbol{S}\|}{\|\boldsymbol{X}\|} = \frac{\sum_{i=1}^{p} w_{ik^*}^t x_i}{\sum_{i=1}^{p} x_i} \tag{5.30}$$

并且需大于或等于设定的警戒门槛值,才会更新记忆,即更新此优胜神经元的连接权重值。若无法找到通过检验的神经元,表示其输入样型与目前记忆不够相似,需另行设立竞争层的其他输出神经元以代表不同群组。

5.5.2 ART1 网络算法

ART1 以警戒值测试来解决稳定性与可塑性间的矛盾。由于能够通过测试的输出层处理单元可能不只一个,故以相似度为评级基准,对相似度最高到最低的输出层处理单元,逐一进行警戒值测试,其须能抗拒外界的干扰,自我更新旧有记忆,并具有足够的可塑性让网络可以快速学习及纳入新的记忆。

ART1 网络的训练过程分为 8 个步骤,如图 5.14 所示(Rao & Rao 1995;Freeman & Skapura,1991):

步骤 1:网络参数的基本条件限制。

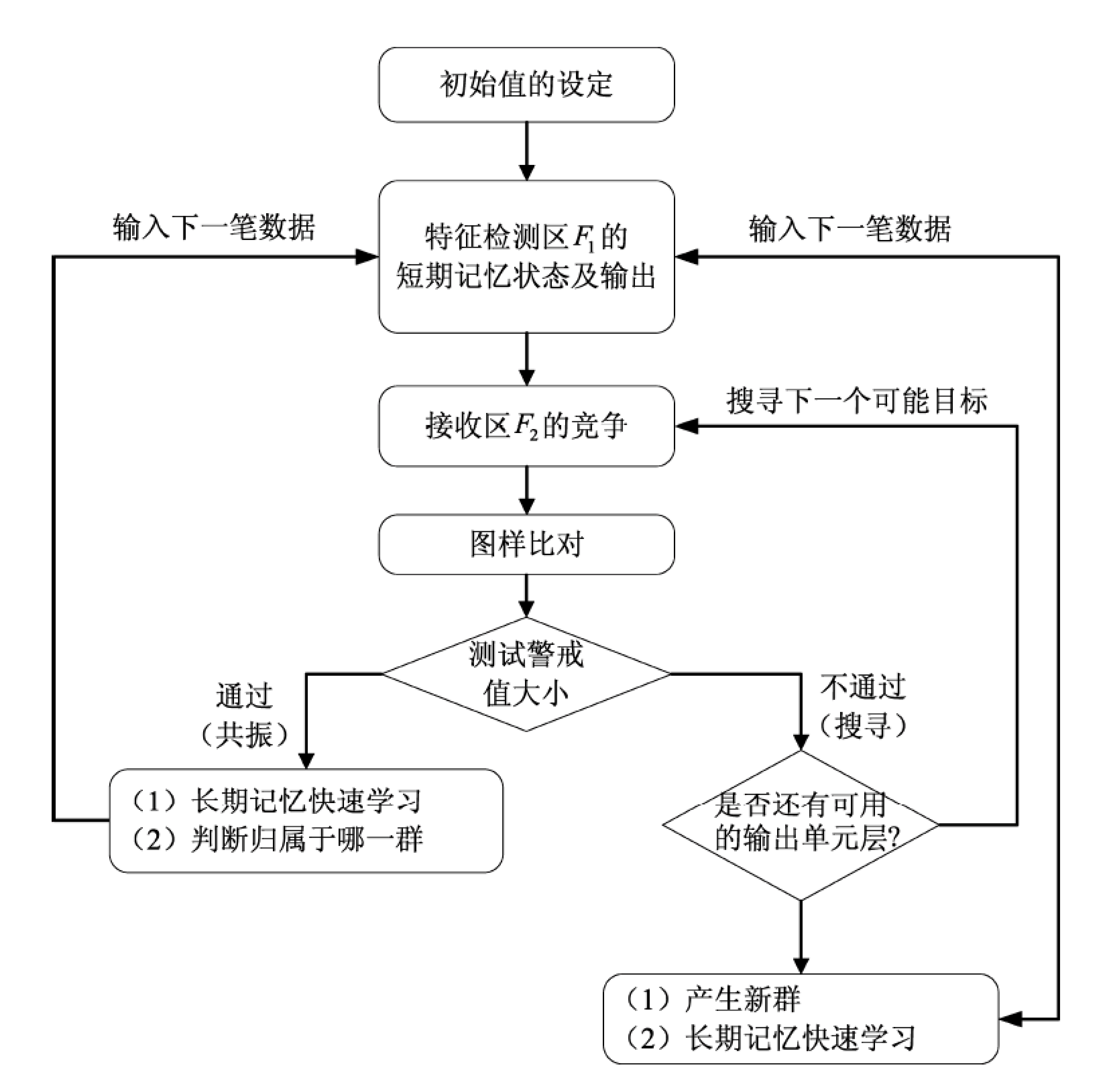


图 5.14 ART1 算法流程图

假设有 p 个输入神经元，$i=1,\cdots,p$，q 个输出神经元的 ART1 神经网络，$k=1,2,\cdots,q$。A、B、C、D 为大于 0 的正值，其中，$\max\{D,1\}<B<D+1$。L 为大于 1 的参数，警戒门槛值 ρ 为介于 0～1 的任意实数，$0<\rho\leqslant 1$。

步骤 2：初始参数的设定。

开始时 ART1 尚未储存任何图样，特征检测区 F_1 神经元的初始状态被设定为 $x_i=\dfrac{-B}{1+C}$，由 F_1 到 F_2 的初始连接权重值 $w_{ki}^b=\lambda_i$，其权重值范围 $0<\lambda_p<\cdots<\lambda_2<\lambda_1<\dfrac{L}{L-1+p}$，由 F_2 到 F_1 的初始连接值 $w_{ik}^t>\dfrac{B-1}{D}$。

步骤 3：输入训练数据与输入向量 $\boldsymbol{I}$，计算特征检测区 F_1 的短期记忆(short-term memory，STM)状态及输出。当图样向量 $\boldsymbol{I}$ 输入至 F_1 后，F_1 的短期记忆状态会变为

$$x_i=\frac{I_i}{1+A(I_i+B)+C},\quad i=1,\cdots,p\text{，而 }F_1\text{ 的输出为 }s_i=\begin{cases}1, & x_i>0\\0, & x_i\leqslant 0\end{cases}\text{。}$$

步骤 4：接收区 F_2 的竞争。

将 F_1 的信号送到接收区 F_2，则接收区 F_2 的输入为 $y_k=\sum_{i=1}^{p}s_i\times w_{ki}^b$，$k=1,2,\cdots,q$，在经过竞争之后，在接收区 F_2 中只有一个神经元 k^* 会因为获胜而送出输出信号，即

$$u_k=\begin{cases}1, & y_{k^*}=\max\limits_k(y_k)\\0, & \text{其他}\end{cases}\text{。}$$

步骤 5：训练数据相似度比对。

当接收区 F_2 的输出信号传到特征检测区 F_1 之后，可以得到样板信号：$v_i = \sum_{k=1}^{q} u_k w_{ik}^t = u_{k^*} w_{ik^*}^t$，此时，输入训练数据会和样板信号发生作用，而使得特征检测区 F_1 的短期记忆状态变成 $x'_i = \frac{I_i + Dv_i - B}{1 + A(I_i + Dv_i) + C}$，因此可重新计算特征检测区 F_1 的输出值 $s'_i = \begin{cases} 1, & x'_i > 0 \\ 0, & x_i \leqslant 0 \end{cases}$。

步骤 6：警戒门槛值检验。

决定输出图样和样板信号的匹配程度 $\frac{|\boldsymbol{S}|}{|\boldsymbol{I}|} = \sum_{i=1}^{p} s'_i / \sum_{i=1}^{p} I_i$。若 $\frac{|\boldsymbol{S}|}{|\boldsymbol{I}|} < \rho$，则刚才获胜的神经元 k^* 会被重置，将使得其输入 y_{k^*}，一直被设定为 0，直到新的图样输入为止。此时，特征检测区 F_1 的短期记忆状态及输出值均改回步骤 2 的值，回到步骤 3，以搜寻下一个可能的目标。若 $\frac{|\boldsymbol{S}|}{|\boldsymbol{I}|} \geqslant \rho$，则表示短期记忆已经进入共振状态，此时将开始学习长期记忆（long-term memory，LTM）。

步骤 7：长期记忆快速学习。

在搜寻结束之后，只有最后获胜之神经元 k^*，其连接值能被改变，而该连接值将被更新为 $w_{k^*i}^b = \frac{L \times s'_i}{L - 1 + |\boldsymbol{S}|}$ 和 $w_{ik^*}^t = s'_i, i = 1, 2, \cdots, p$。

步骤 8：输入新的训练数据，回到步骤 2。

5.5.3 适应性共振网络范例

本节以 4 个 2×3 的输入图像为例，如表 5.4 所示，说明 ART1 算法计算过程。$p=6$，首先将这 4 个图像转换成 6×1，以 1 代表■，0 代表□，因此得到 4 组输入向量集，分别为 $\boldsymbol{I}^1=(1,1,1,0,0,0)$、$\boldsymbol{I}^2=(0,0,0,1,1,1)$、$\boldsymbol{I}^3=(1,0,0,0,1,1)$、$\boldsymbol{I}^4=(1,1,1,0,1,0)$，演算过程如下所示。

表 5.4　ART1 示例的四种输入图像

1	2	3	4
■■■ □□□	□□□ ■■■	■□□ □■■	■■■ □■□

步骤 1：初始参数设定，$A=2.0$，$B=1.5$，$C=4.0$，$D=0.6$，$L=2.0$，$\rho=0.5$。

步骤 2：设定 $0 < w_{ki}^b < \frac{2}{2-1+6}$，所以 $\boldsymbol{W}_1^b = (1/7, 1/7, 1/7, 1/7, 1/7, 1/7)$，设定 $w_{ik}^t > \frac{1.5-1}{0.6}$，所以 $\boldsymbol{W}_1^t = (1,1,1,1,1,1)$。

步骤 3：输入第一组训练样本向量 $\boldsymbol{I}^1 = (1,1,1,0,0,0)$，

计算特征检测区 F_1 的状态向量 $\boldsymbol{X}$，$x_i = \frac{I_i}{1+2(I_i+1.5)+4}$，所以 $\boldsymbol{X} = (1/10, 1/10, 1/10, 0, 0, 0)$，$\boldsymbol{S} = (1,1,1,0,0,0)$。

步骤 4：计算接收区 F_2 的输入向量为 $\boldsymbol{Y}$，此时网络接收区 F_2 只有一个神经元，其相似度为 $y_1 = \sum_{i=1}^{p} s_i \times w_{1i}^b = 3/7$。

步骤 5：接收区 F_2 至特征检测区 F_1 的输入信号 $v_i = u_1 \times w_{i1}^t$，$\boldsymbol{V} = (1,1,1,1,1,1)$，此时特征检测区 F_1 状态变成 $x_i' = \dfrac{I_i + Dv_i - B}{1 + A(I_i + Dv_i) + C}$。

$$x_1' = x_2' = x_3' = \frac{1 + 0.6 \times 1 - 1.5}{1 + 2(1 + 0.6 \times 1) + 4} = 0.0098$$

$$x_4' = x_5' = x_6' = \frac{0 + 0.6 \times 1 - 1.5}{1 + 2(0 + 0.6 \times 1) + 4} = -0.1216$$

所以重新计算特征检测区 F_1 的输出值 $\boldsymbol{S}' = (1,1,1,0,0,0)$。

步骤 6：警戒门槛值检验，$\dfrac{|\boldsymbol{S}'|}{|\boldsymbol{I}|} = 3/3 = 1.0$，大于警戒门槛值 $\rho = 0.5$。

步骤 7：更新连接权重，$\boldsymbol{W}_1^b = (0.5,0.5,0.5,0,0,0)$，$\boldsymbol{W}_1^t = (1,1,1,0,0,0)$。

步骤 8：输入第 2 笔训练数据 $\boldsymbol{I}^2 = (0,0,0,1,1,1)$，回到步骤 2。

步骤 2～7：以下为简化说明，仅列出计算结果与权重更新结果，$\boldsymbol{S} = (0,0,0,1,1,1)$，

$$\boldsymbol{V} = (1,1,1,0,0,0),\ x'_1 = x'_2 = x'_3 = \frac{0 + 0.6 \times 1 - 1.5}{1 + 2(0 + 0.6 \times 1) + 4} = -0.1216,$$

$$x'_4 = x'_5 = x'_6 = \frac{1 + 0.6 \times 0 - 1.5}{1 + 2(1 + 0.6 \times 0) + 4} = -0.0714,$$

所以 $\boldsymbol{S}' = (0,0,0,0,0,0)$，$\dfrac{|\boldsymbol{S}|}{|\boldsymbol{I}|} = 0/3 = 0$，小于警戒门槛值，且无其他输出层神经元可供警戒值检验，因此产生第二个输出层神经元，且 $w_{2i}^b = \dfrac{I_i}{0.5 + \|\boldsymbol{I}\|}$，$w_{2i}^t = I_i$，所以 F_1 与 F_2 之间连接权重为 $\boldsymbol{W}_2^b = (0,0,0,2/7,2/7,2/7)$，$\boldsymbol{W}_2^t = (0,0,0,1,1,1)$。

步骤 8：输入第 3 笔训练数据 $\boldsymbol{I}^3 = (1,0,0,0,1,1)$，回到步骤 2。

步骤 2～7：$\boldsymbol{S} = (1,0,0,0,1,1)$，网络有两个输出神经元，第一个输出神经元的相似度 $y_1 = \sum_{i=1}^{p} s_i \times w_{1i}^b = 1/2$，第一个输出神经元的相似度 $y_2 = \sum_{i=1}^{p} s_i \times w_{2i}^b = 4/7$，所以第二个神经元胜出，并进行警戒值测试，$v_i = u_2 w_{i2}^t$，$\boldsymbol{V} = (0,0,0,1,1,1)$，$\boldsymbol{S}' = (0,0,0,0,1,1)$，$\dfrac{|\boldsymbol{S}'|}{|\boldsymbol{I}|} = 2/3 = 0.67$，大于警戒门槛值，因此更新 $\boldsymbol{W}_2^b = (0,0,0,0,2/3,2/3)$，$\boldsymbol{W}_2^t = (0,0,0,0,1,1)$。

步骤 8：输入第 4 笔训练数据 $\boldsymbol{I}^4 = (1,1,1,0,1,0)$，回到步骤 2。

步骤 2～7：$\boldsymbol{S} = (1,1,1,0,1,0)$，第一个输出神经元的相似度 $y_1 = \sum_{i=1}^{p} s_i \times w_{1i}^b = 3/2$，第一个输出神经元的相似度 $y_2 = \sum_{i=1}^{p} s_i \times w_{2i}^b = 2/3$，所以第一个神经元胜出，并进行警戒值测试，$v_i = u_1 w_{i1}^t$，$\boldsymbol{V} = (1,1,1,0,0,0)$，$\boldsymbol{S}' = (1,1,1,0,0,0)$，$\dfrac{|\boldsymbol{S}'|}{|\boldsymbol{I}|} = 3/3 = 1.0$，大于警戒门槛值，因此更新 $\boldsymbol{W}_1^b = (0.5,0.5,0.5,0,0,0)$，$\boldsymbol{W}_1^t = (1,1,1,0,0,0)$。

步骤 8：所有训练数据均输入，停止网络训练。可得两组分群结果，如表 5.5 所示。其

中,“图形 1 与图形 4 为相似图形;图形 2 与图形 3 是相似图形”。

表 5.5 ART1 示例的四种输入图像的分群结果

群组	样本	
#1 ■ ■ ■ □ □ □	1 ■ ■ ■ □ □ □	4 ■ ■ ■ □ ■ □
#2 □ □ □ □ ■ ■	2 □ □ □ ■ ■ ■	3 ■ □ □ □ ■ ■

5.6 R 语言与人工神经网络

5.6.1 反向传播人工神经网络

本节利用皮马族印第安人糖尿病数据集建构一个反向传播人工神经网络模型,用以预测是否会罹患糖尿病。可以通过 R 的扩充套件 **RSNNS**(Bergmeir & Benítez,2012)进行反向传播人工神经网络的模型构建。

首先,选择 200 笔的训练数据,并随机切割 10%的数据作为测试,目的是避免人工神经网络的训练过程中有过度配适的情况,而此数据的目标属性(type)为一个二分类变量,代表在人工神经网络架构中的输出层需有两个神经元,因此需将之重新编码转换为一组指针变量(indicator variable)。此外,为避免 7 个属性间不同尺度影响分析结果,亦需进行数据标准化。

```
library(MASS)
data(Pima.tr)
set.seed(1111)                                   #设定随机种子
#将数据顺序重新排列
Pima.tr <- Pima.tr[sample(1:nrow(Pima.tr),length(1:nrow(Pima.tr))),]
PimaValues <- Pima.tr[,1:7]
PimaTargets <- decodeClassLabels(Pima.tr[,8])    #目标属性重新编码
Pima.tr <- splitForTrainingAndTest(PimaValues,PimaTargets,ratio=0.1)
Pima.tr <- normTrainingAndTestSet(Pima.tr)
```

完成数据切割后,接着以 **mlp** 函数训练反向传播人工神经网络模型。在此函数中,指定隐藏层神经元个数为 14、学习率为 0.01、最大迭代次数 100 为停止条件。训练完成的模型可通过 **plotIterativeError** 函数功能了解模型的误差收敛情况,而 **weightMatrix** 函数则可用以提取模型中各神经元连接上的权重。

```
model <- mlp(Pima.tr$inputsTrain,Pima.tr$targetsTrain,
  size=14,learnFuncParams=0.01,maxit=100,inputsTest=Pima.tr$inputsTest,
  targetsTest=Pima.tr$targetsTest)
#size:隐藏层神经元个数
#learnFuncParams:学习率
```

```
#maxit:最大迭代次数
plotIterativeError(model)
weightMatrix(model)
```

图 5.15 即为反向传播类神经网络的误差收敛图,纵轴为残差平方和(sum of square error,SSE),横轴为迭代次数。从中可看出此模型误差在迭代次数大于 60 之后便趋于稳定。然而,不同的参数设定对于人工神经网络模型的影响甚大,故可以通过尝试错误法找出最佳的参数组合。

```
p_table=expand.grid(size=c(12,13,14,15,16),learning.rate=c(0.001,0.01,0.1))
for (i in 1:nrow(p_table)){
model <-mlp(Pima.tr$inputsTrain,Pima.tr$targetsTrain,size=p_table[i,1],learnFuncParams=p_table[i,
2],
  maxit=100,inputsTest=Pima.tr$inputsTest,targetsTest=Pima.tr$targetsTest)
p_table$TestError[i]=model$IterativeTestError[100]
}
p_table
```

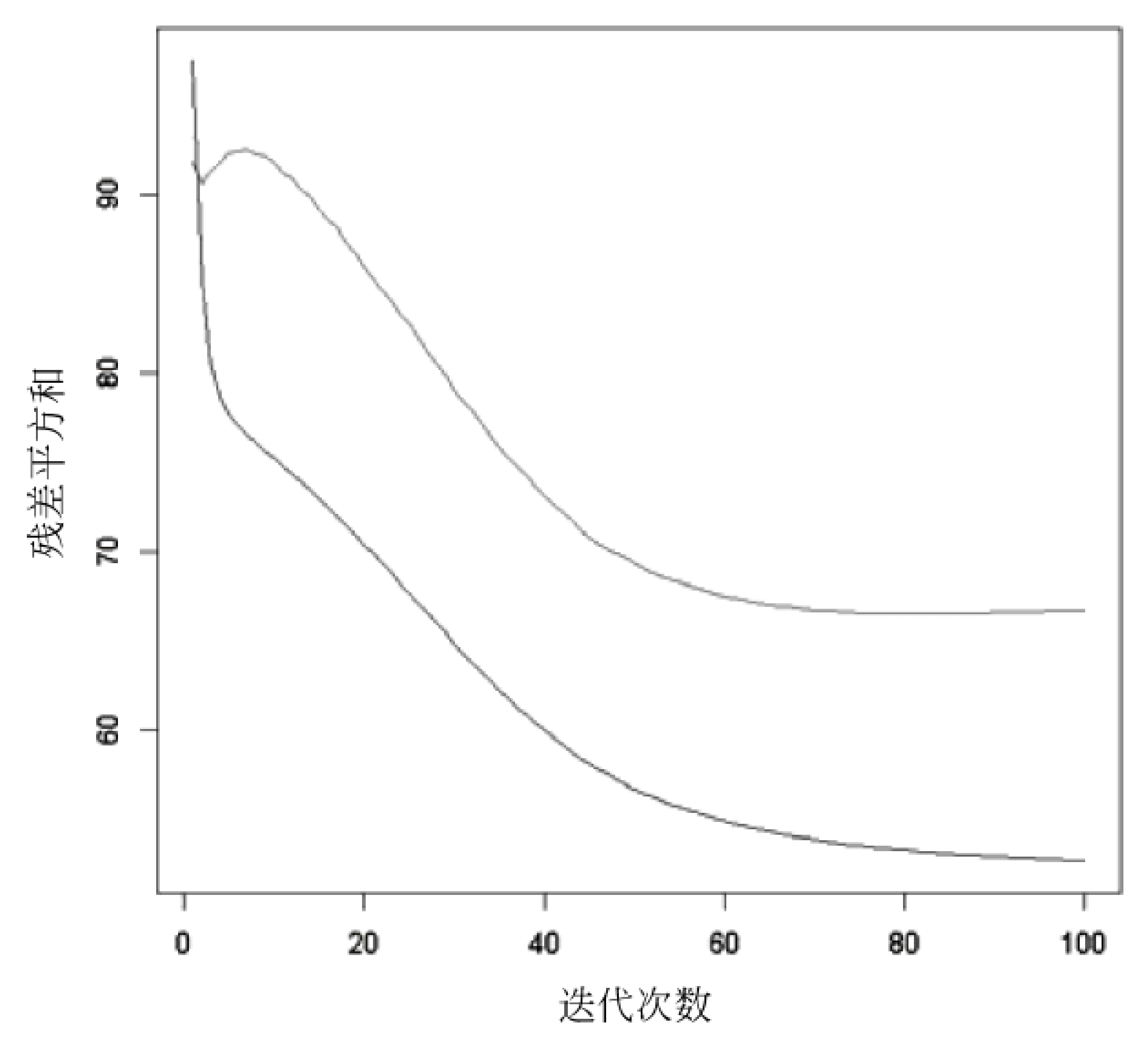


图 5.15　反向传播人工神经网络误差收敛图

表 5.6 即为设定 5 水平的隐藏层神经元个数与 3 水平的学习率进行交叉比较,而在这 15 种组合中,隐藏层神经元个数为 13、学习率为 0.01 的组合测试数据的误差最小,因此可用此参数组合的反向传播人工神经网络模型进行另一组 332 笔数据的预测。通过分类矩阵可计算分类正确率为 0.804。

```
Pima.te[,1:7] <-normalizeData(Pima.te[,1:7])
predictions <-predict(model,Pima.te[,1:7])
table <-confusionMatrix(Pima.te[,8],predictions)
accuracy=sum(diag(table))/sum(table);accuracy
```

表 5.6 反向传播人工神经网络参数设定误差比较

隐藏层神经元个数	学习率		
	0.001	0.01	0.1
12	10.05	7.46	9.41
13	9.71	7.24	9.30
14	10.06	7.41	9.14
15	9.96	7.44	9.93
16	9.77	7.37	9.26

5.6.2 自组织映射网络

本节则以皮马族印第安人糖尿病数据集中前 7 个连续型的属性构建自组织映射网络以将 532 笔数据进行分群。可以通过 R 的扩充套件 **kohonen**(Wehrens & Buydens,2007)构建自组织映射网络模型。首先,从扩充套件 **MASS**(Venables & Ripley,2002)中加载数据集,同时将属性标准化避免不同尺度影响分群结果。

```
library(MASS)
data("Pima.tr")
Pima_class <- rbind(Pima.tr,Pima.te)[,8]
Pima <- scale(rbind(Pima.tr,Pima.te)[,-8])
```

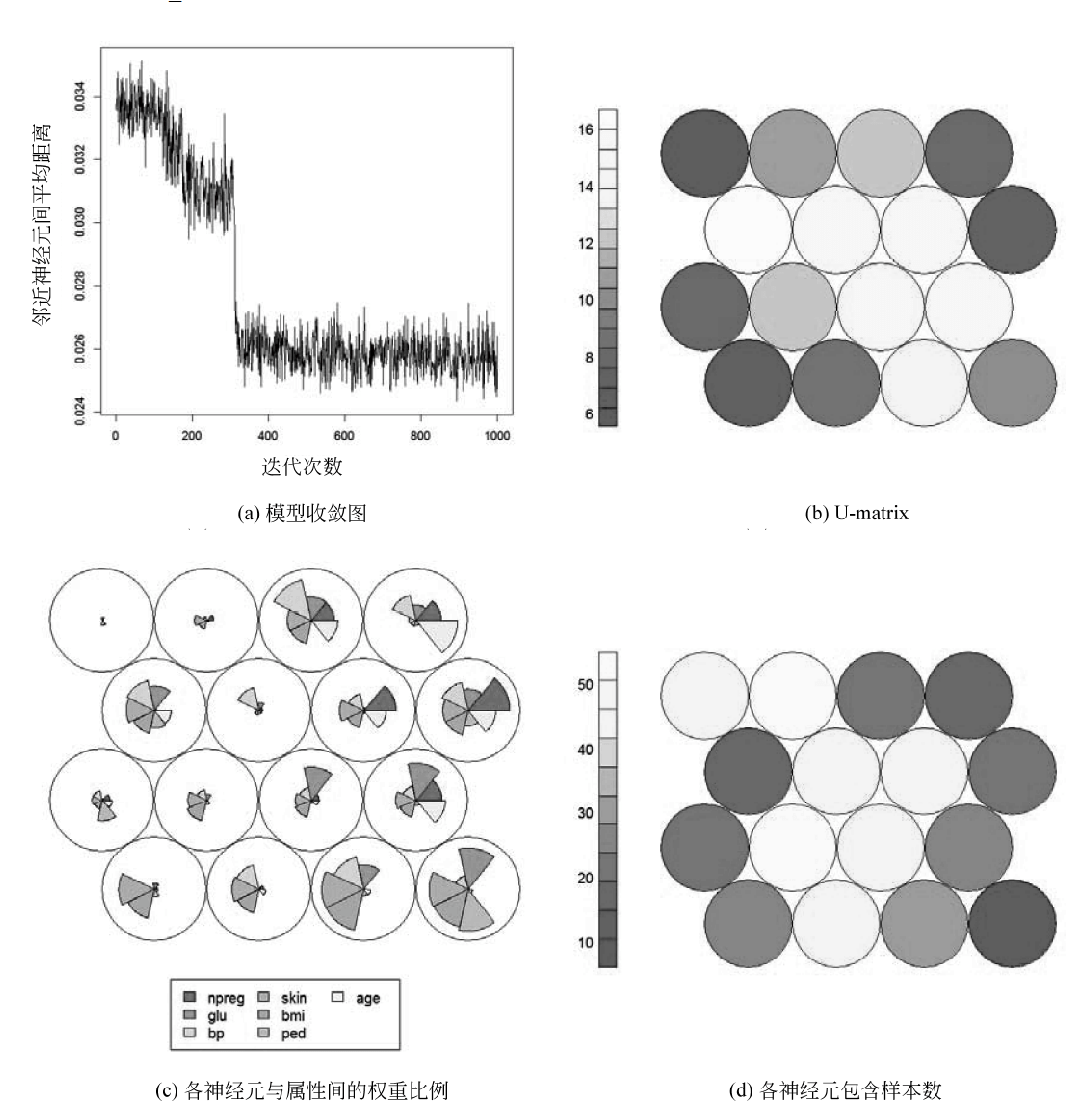
接着,通过 **som** 函数建立模型。在此,指定输出层为 4×4 的六角形网络拓扑结构,并设定最大迭代次数为 1000 次、学习率为从 0.05 递减至 0.01。完成训练后,可进一步通过 **plot** 函数检查模型的收敛情况。

```
library(kohonen)
set.seed(1111)
Pima_som <- som(data= Pima,grid= somgrid(4,4,"hexagonal"),
  rlen= 1000,alpha= c(0.05,0.01))
#grid可设定输出层大小,"hexagonal"代表六角形网络拓扑结构
##"rectangular"代表正方形网络拓扑结构
#rlen为最大迭代次数
#alpha为学习率,两个数字分别为变化前起始值与变化后结束值
plot(Pima_som,type= "changes")
```

如图 5.16 (a)所示,模型经过 1000 次的迭代已趋近稳定收敛的情况,图 5.16(b)呈现整体的网络拓扑结构(套件中又称 U-matrix),邻近神经元间的颜色越接近代表相似度越高,可凝聚为一群;反之,若颜色差异甚大,代表可视之为不同群。图 5.16(c)显示各输出神经元与输入属性间的权重比例,可用以了解分群之特性,图 5.16(d)则呈现各输出神经元所包含的样本数。

```
plot(Pima_som,type= "dist.neighbours")
plot(Pima_som,type= "codes")
```

```
plot(Pima_som,type="counts")
```



(a) 模型收敛图　　(b) U-matrix

(c) 各神经元与属性间的权重比例　　(d) 各神经元包含样本数

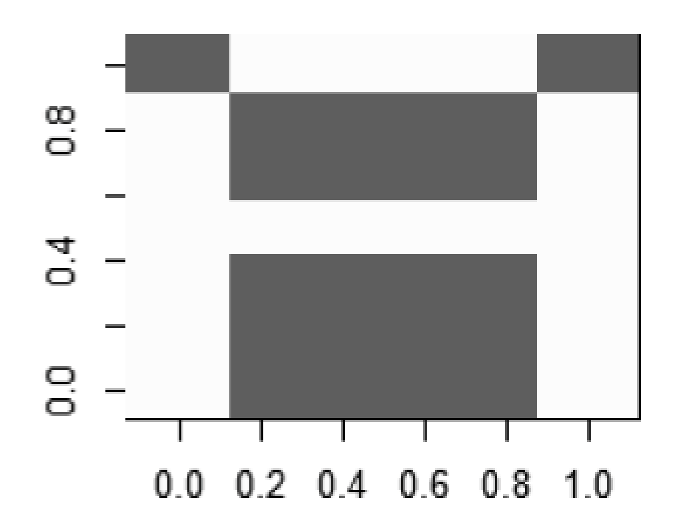
图 5.16　自组织映射网络主要输出图形

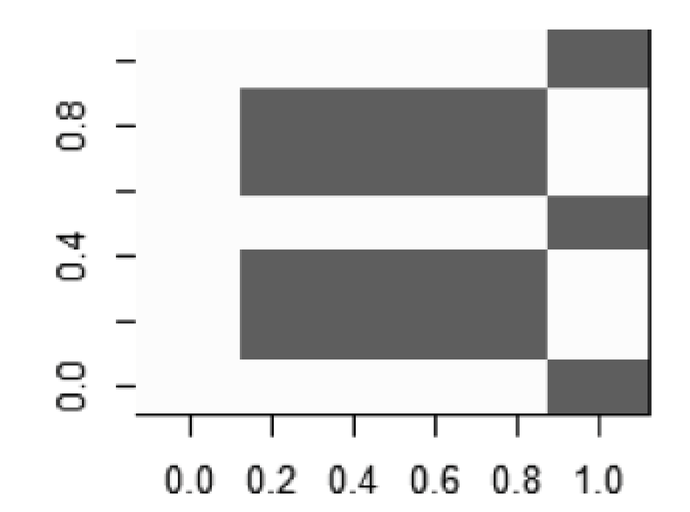
5.6.3　自适应共振理论人工神经网络

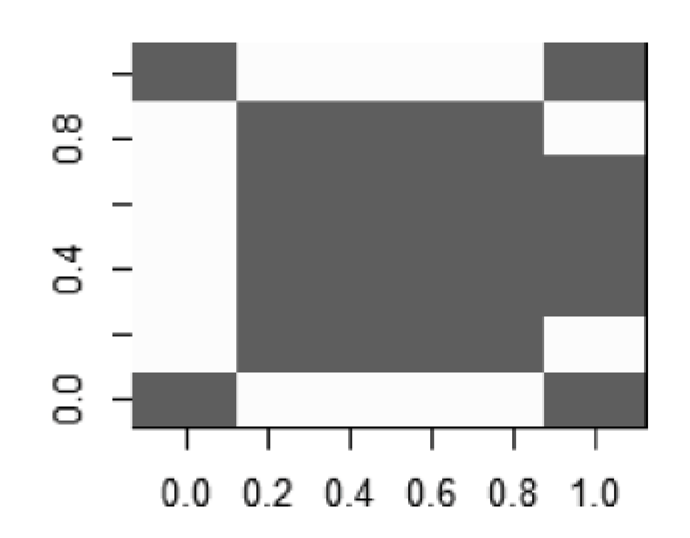
在本节以扩充套件 **RSNNS**(Bergmeir & Benítez,2013)中的一组范例数据说明如何运用 **art1** 函数构建自适应共振理论人工神经网络以进行样型分群。此组数据包含 26 笔 7×5 的二维图形数据,每一个图形均由 0 或 1 的二元数值构成。图 5.17 为此组数据的前 9 笔图形,红色为 0,米黄色为 1。

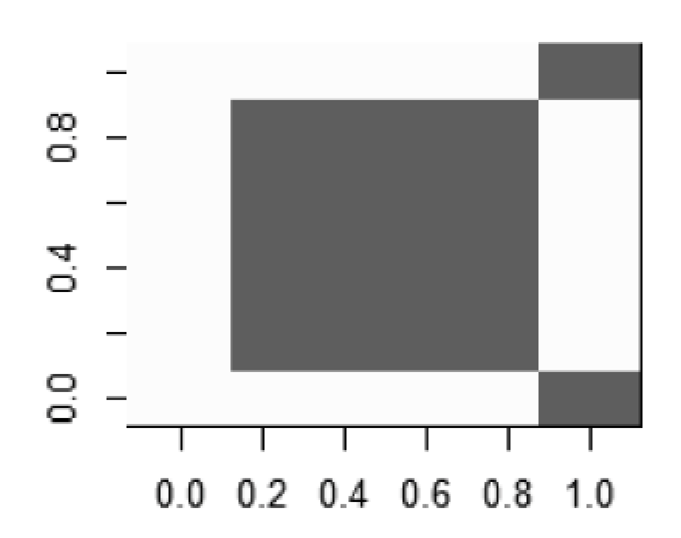
```
library(RSNNS)
data(snnsData)
patterns <- snnsData$art1_letters.pat
```

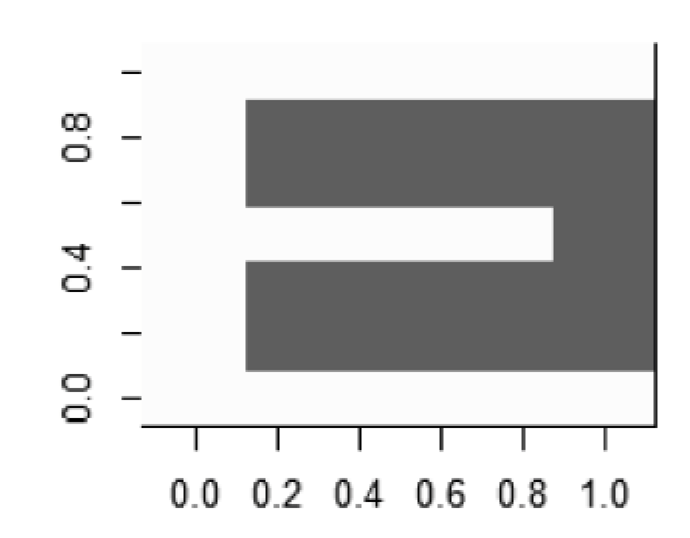
```
inputMaps <- matrixToActMapList(patterns,nrow=7)
par(mfrow=c(3,3))
for (i in 1:9) plotActMap(inputMaps[[i]])
```

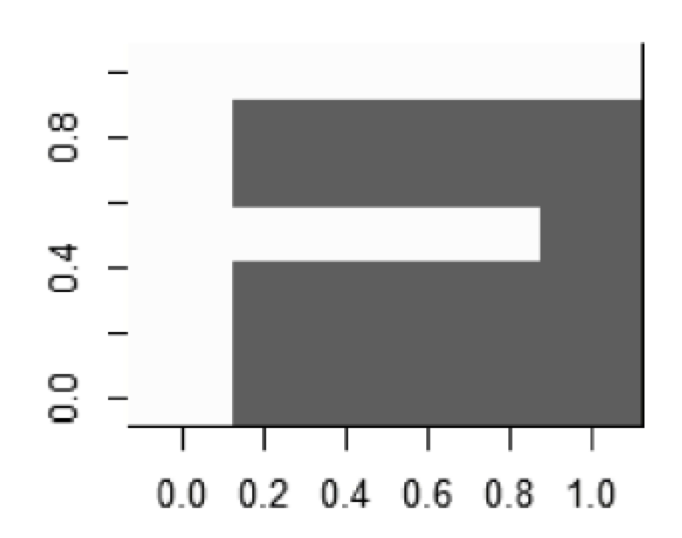


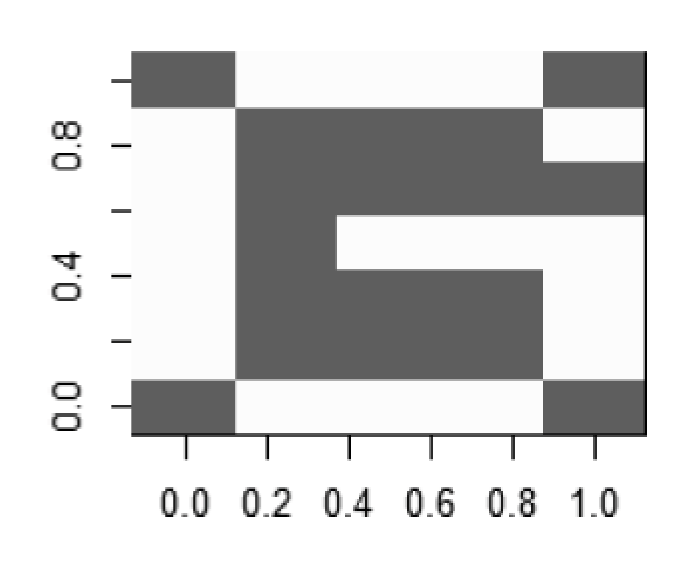


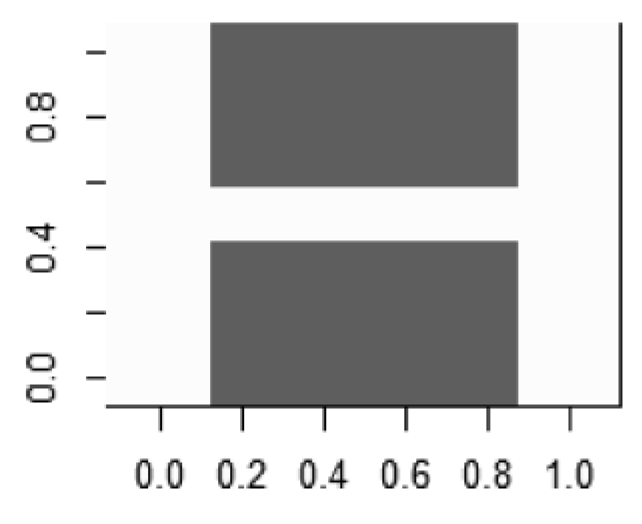












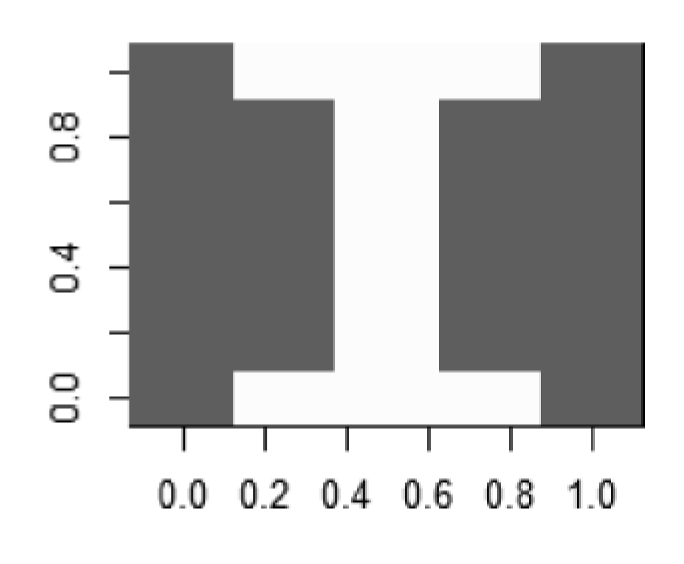


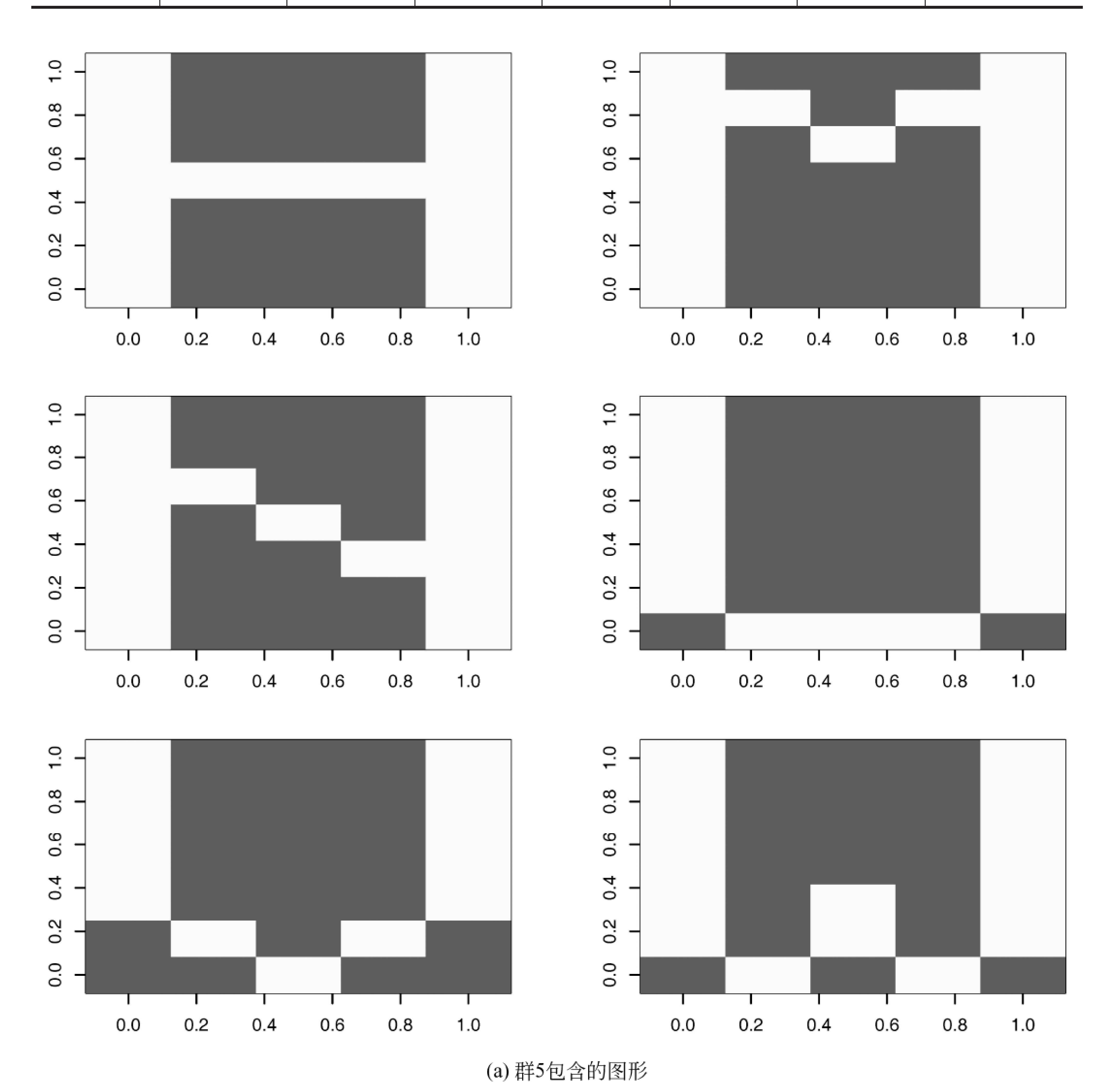
图 5.17　ART1 部分输入图形数据

接着，给定警戒值参数为 0.5，最大迭代次数为 100，建立 ART1 分群模式，其分群结果如表 5.7 所示，共分成 7 群，其中，第 7 群包含的样本数最多，第 5 群次之，第 5 群与第 7 群所包含的图形如图 5.18(a)与图 5.18(b)所示。

```
model <- art1(patterns,dimX=7,dimY=5,learnFuncParams=c(0.5,0,0),maxit=100)
#learnFuncParams 为学习率
#maxit 为最大迭代次数
table(encodeClassLabels(model$fitted.values))
```

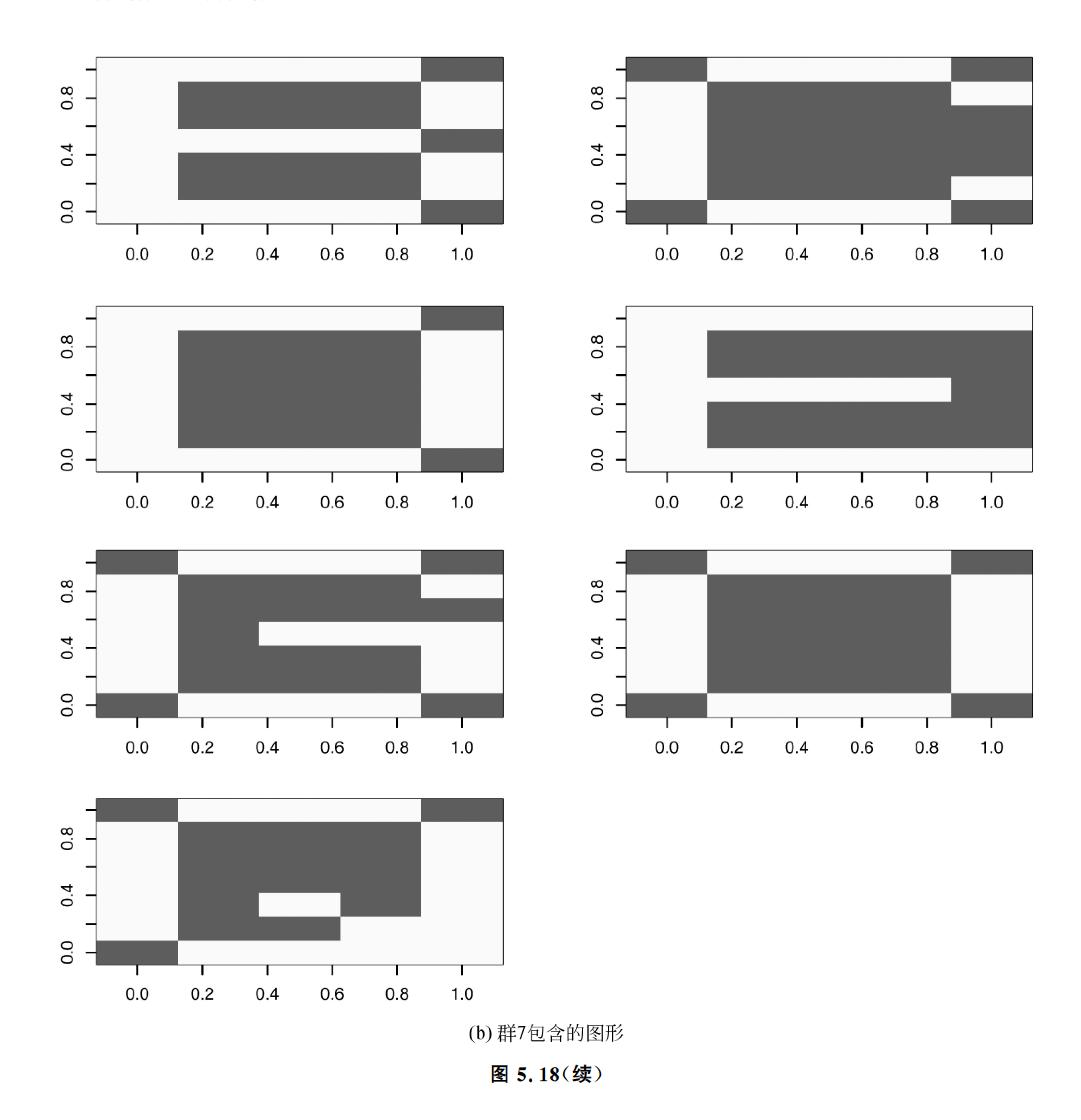
表 5.7　ART1 给定警戒值为 0.5 的各分群样本数

分群	群 1	群 2	群 3	群 4	群 5	群 6	群 7
样本数	3	3	2	3	6	2	7



(a) 群5包含的图形

图 5.18　ART1 给定警戒值为 0.5 的分群结果



(b) 群7包含的图形

图 5.18(续)

5.7 应用实例——半导体生产周期时间预测与管控

5.7.1 案例简介

半导体制造受到工件回流、动态到达、生产流程长与瓶颈机台飘移等限制条件与不确定性的影响,使得制品水位的生产周期时间与产出变得难以精确预测(Kuo *et al.*,2011)。本案例(Chien *et al.*,2012)以生产线搜集的制造数据,考虑领域知识以推演实证规则,借由控制输入因子以达成周期时间与产出的控管,并整合不同的数据挖掘技术,包括自组织映射网络、多项式回归(polynomial regression,PR)分析法与反向传播人工神经网络等,构建生产周期时间预测模式,并以某半导体制造厂商为实证案例,检验研究效度。

5.7.2　数据分群

由于制造数据容易受到人为管理因素的干扰，例如，生产投入量的改变，或是产品批次优先级的调整，因此取得数据后必须先做数据准备再进行后续分析。首先删除不合适的部分并保留合适的数据群组，分为训练数据组与测试数据组，前者应用于模式建立时的输入数据，后者应用于检验该建立模式的信度及效度；并以自组织映射网络将数据分群，再以决策树进行分类规则的提取；接续则构建一个多项式回归模型来描绘在制品(work-in-process，WIP)水平与作业数(Move)及 WIP 水平与周期时间(cycle time，CT)间的关系。

1. 数据准备

本案例搜集相关属性与数据，包括 WIP、Move、CT、产能(Capacity)以及利用率(Utilization)等。由于半导体制造自动累积巨量数据，因此可将属于同一时间区隔的数据点合并为同一组数据集，以减少数据库储存空间，并加速模式构建的效率。然而，若遇遗漏值时，应检查可否采用数量化的方式还原，倘若发现某笔数据的主要属性均有遗漏值时，则应移除该笔数据，避免影响模式效度。

此外，利用数据转换方式将各属性下不同衡量尺度的数据值标准化，确保数据的适切性。例如，当产能扩充的同时，WIP 水平与 Move 也会相对提升，且不同产能水平下的相同的 WIP 水平与 Move 代表不同的意义，因此应以各厂的产能水平为基础将所搜集的 WIP 与 Move 数据标准化，如式(5.31)所示：

$$\begin{cases} x_m^* = \dfrac{x_m}{c_m}, & m = 1,2,\cdots,N_0 \\ y_m^* = \dfrac{y_m}{c_m}, & m = 1,2,\cdots,N_0 \end{cases} \tag{5.31}$$

其中，x_m^* 代表第 m 个厂所规定的 WIP 水平，y_m^* 代表第 m 个厂所规定的 Move 水平，N_0 则为厂的总数目。

2. 自组织映射网络

本案例采用 SOM 神经网络先将利用率数据进行分群。由于不同的利用率水平会导致半导体制造厂生产形态的差异，因此以分群方式将生产形态进行聚类分割。通过向量量化与向量投影，可将数据聚类现象绘成拓扑图，利用此图可了解数据点在图上的分布，并以颜色来区分各群。此外，更可以良率分布拓扑图来检查分群的良率表现，进一步探讨各分群间的关联性，亦可找出哪些参数对于分群与良率表现有较大的贡献。完成分群后，接着以决策树进行特征提取与分类规则的描述。经由 SOM 算法进行利用率数据集的分群后，可得最佳的分群群数为 3，如图 5.19，表示以 3 个群组来分隔利用率数据能使得后续的分析更有效率。

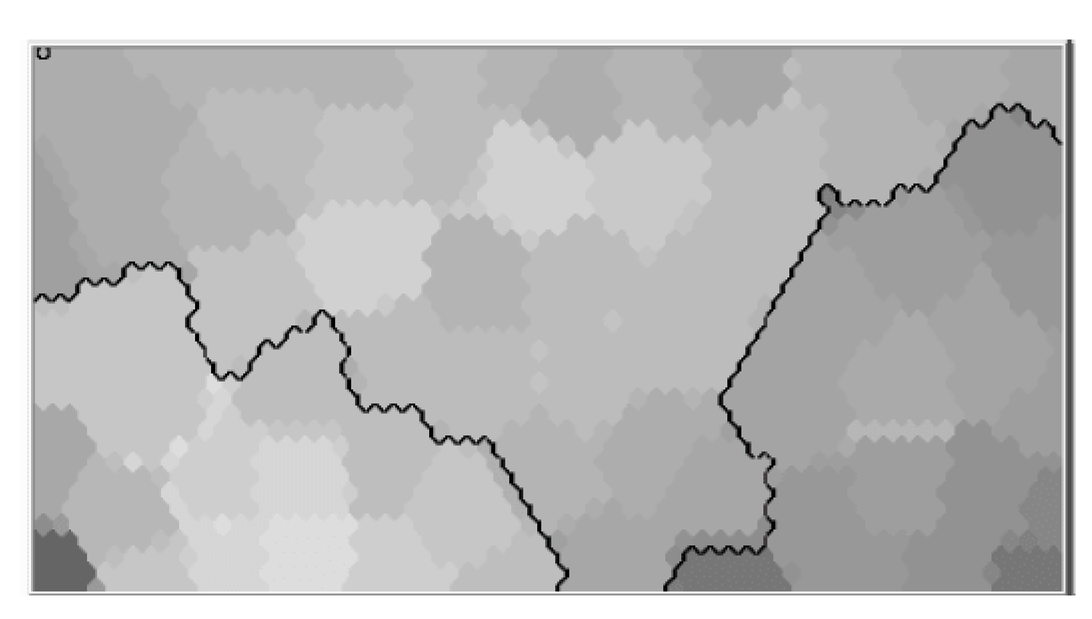


图 5.19　利用率的 SOM 聚类现象的拓扑图

3. 决策树

通过前述的 SOM 分群法取得最适的群数后，本案例采用决策树中卡方自动交互检测(CHAID)算法，找出区隔的标准，并将利用率分为数段区间。在衡量决策树分类规则时，选择以置信度代表此分类节点的纯度，以准确率代表此节点相对于原有类别个数被正确区隔的比例。换言之，期望找到准确率与置信度高的规则来代表分群特征。接着再以决策树提取分类规则，可得出用于区隔 3 个群组的利用率分割值，分别为 0.8 及 1.4，其分支规则所提取的信息如图 5.20 所示。因此，可了解当利用率低于 0.8(群组一)、介于 0.8 与 1.4 之间(群组二)以及高于 1.4(群组三)的生产形态会有显著不同。

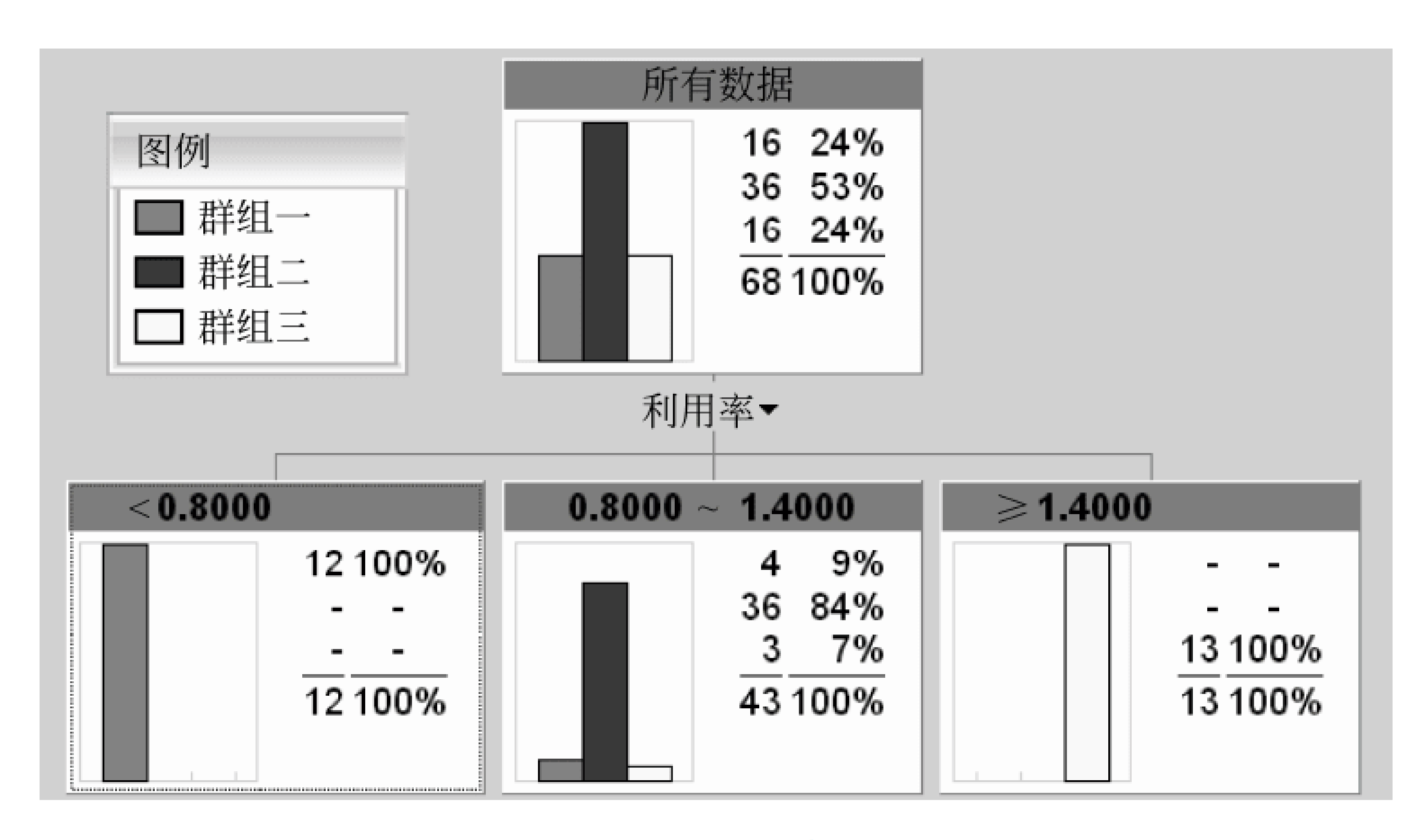


图 5.20 利用率的决策树分类规则

5.7.3 数据配适与预测

完成第一阶段数据的前置处理及提取出数据的分类规则后，第二阶段即可以 WIP、Move、平均流程层数(average layer)以及批次数目(lot)作为构建预测 CT 模式的重要属性。

1. 多项式回归

当 WIP 水平增加时，CT 会以指数形态持续增加，而 Move 则会以相对比例持续增加(简祯富等，2005)。然而，WIP 与 Move 的关系应为正相关，但是当系统产能到达一定的饱和度时，两者之间可能就不是正相关，甚至为负相关。因为 Move 除了和 WIP 相关外，还受到其他因素影响，例如，当 WIP 超过需求还继续增加时，不但对 Move 没有正面帮助，反而会增加现场排货的困难、造成输送带拥塞，以及人员在找货时的困难等，因此在实证数据上反而为负相关。

以数据散布图检查各群组的“WIP 对于 Move”以及“WIP 对于 CT”的相关程度。群组一对于产能限制的敏感度不高，其所贡献的信息极少；群组二及群组三有着极为相似的散布趋势图，因此将群组二与群组三合并后建立多项式回归模型，如图 5.21 所示，其中的 WIP 与 Move 数值均已于阶段一时根据产能限制的水平转化成相对值，并建立多项式回归模型。

2. 反向传播人工神经网络

本案例整合 WIP、平均流程层数以及货批数目建立反向传播人工神经网络模型，以作

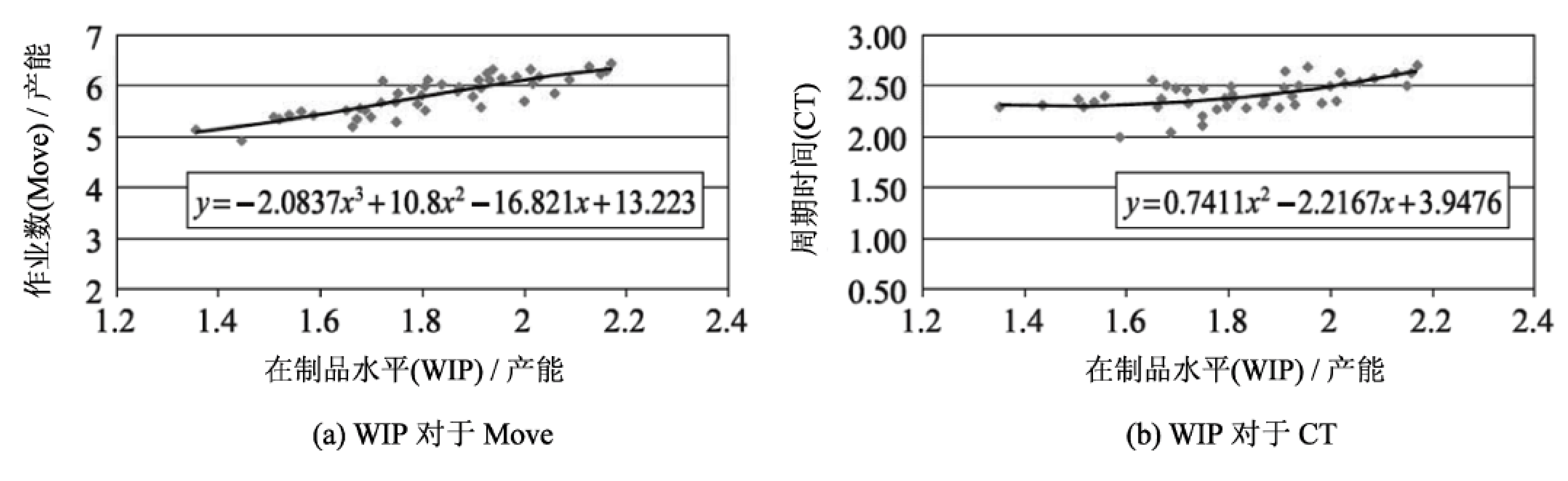


图 5.21　多项式回归模型

为配适多项式回归式中所产生的残差项；与前一步骤的预测模型合并后可得 CT 或 Move 的预测模型，如式(5.32)所示：

$$\hat{\theta} = PR(\text{WIP}) + BPN(\text{average layer}, \text{lot}, \text{WIP}) \tag{5.32}$$

其中，$\hat{\theta}$ 代表 CT 或 Move 的预测值，模式中的前半部为以 WIP 预测 CT 或 Move 的多项式回归模式，后半部则为以平均流程层数、批次数目以及 WIP 预测多项式回归式中残差项的 BPNN 预测模型。

然后，采用 3 个输入节点（WIP、平均流程层数、批次数目）、2 层隐藏层以及 1 个输出节点（CT 预测值）的 BPNN 模型来预测多项式回归中的残差项，图 5.22 为 WIP 对于 Move 所建立的多项式回归预测模式中的残差项序列，因此可利用式(5.32)，找出 CT 时间与 Move 数目的预测模型，而 Move 预测序列值如图 5.23 所示，由误差值仅有 2.4% 可知，所提出的模式的配适结果良好。

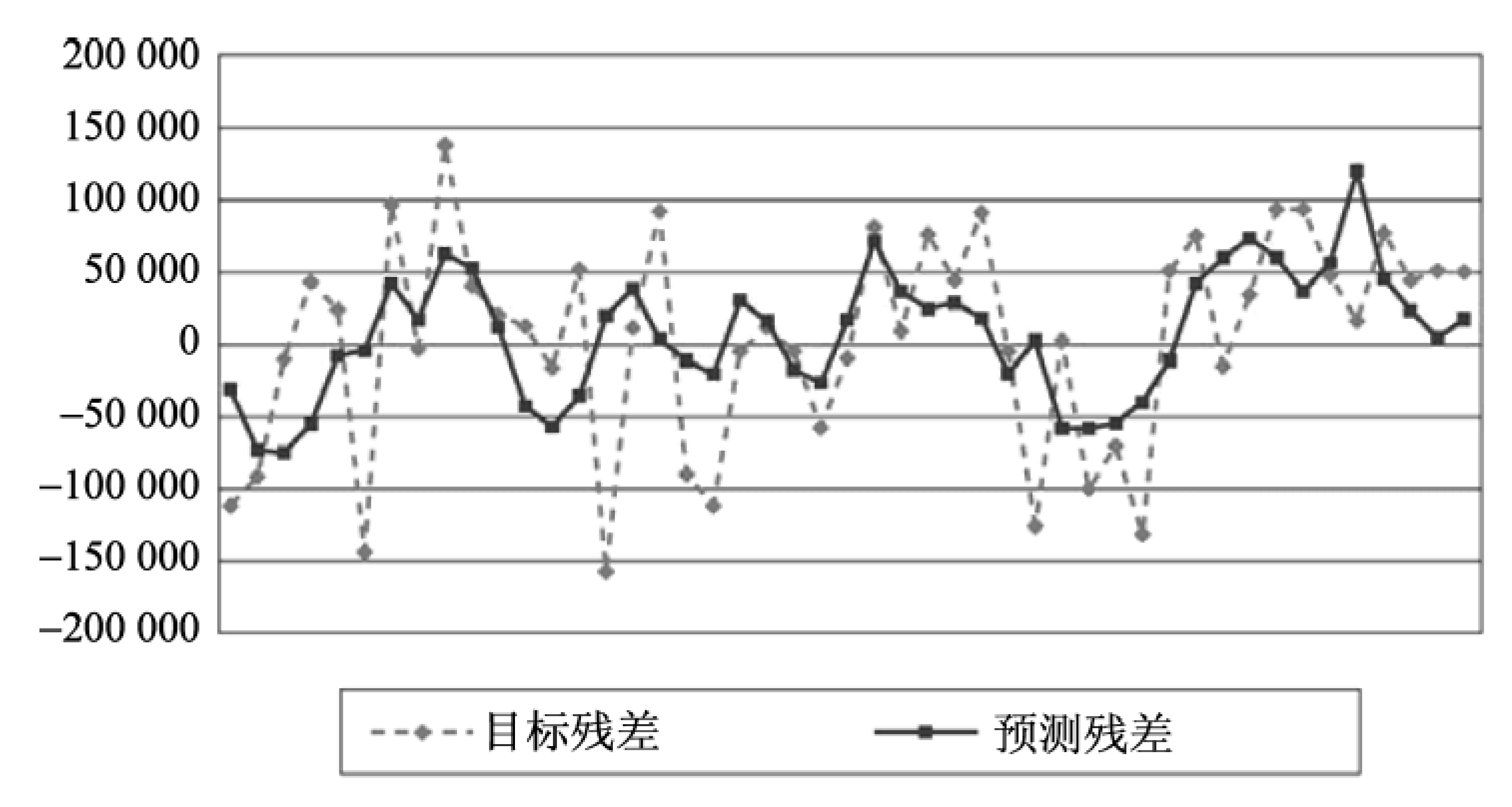


图 5.22　WIP 对于 Move 多项式回归预测模式的残差项序列图

5.7.4　信息整合与敏感度分析

在第二阶段的 CT 或 Move 预测模式的构建中，以 WIP、平均流程层数、批次数目三个属性来预测 CT 或 Move 时间。由于当 WIP 增加时，CT 时间也会增加，因此，在此阶段中，可以与专家讨论后的既有信息、CT 或 Move 的预测模式以及 Move、CT 与 WIP 之间的关系，以规定最适的 WIP 水平。

以第二阶段所取得的模式为基础，计算在不同 WIP 水平下的 Move 数目与 CT 时间预

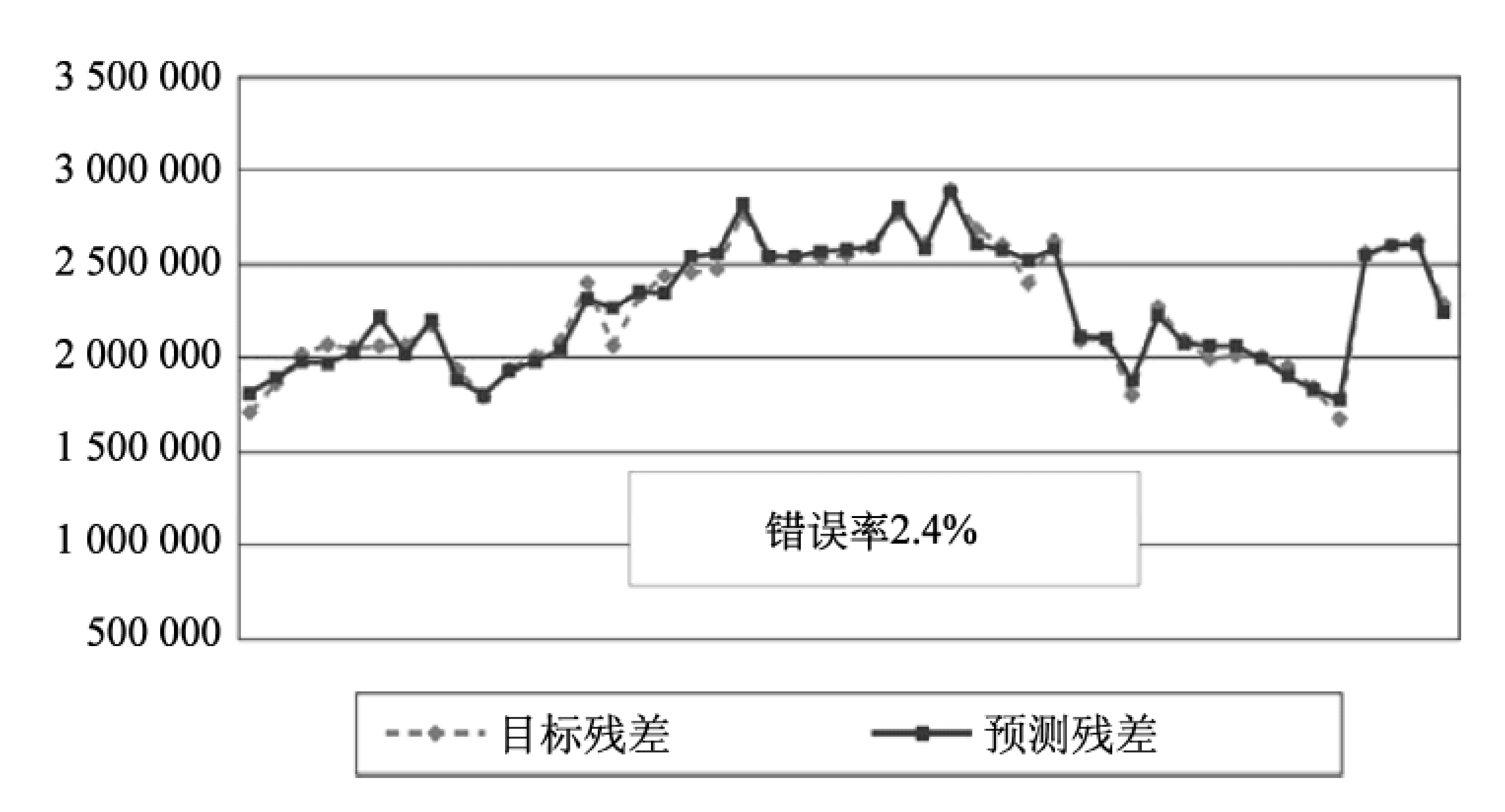


图 5.23　Move 预测值的序列图

测值(此处 WIP 与 Move 均为转换产能限制后的新数据点,分别以 WIP/产能与 Move/产能表示),并将这些数据以敏感度分析(sensitivity analysis)如表 5.8 所示。其中,"slope"代表"WIP 对于 Move"所构建的多项式回归模型的回归参数估计值,当斜率逼近于 0 时,即代表目前所规定的 WIP 水平已达产能上限,此时若再增加 WIP 水平,只会徒然增加 CT,但对于 Move 的产出数目并无正向贡献。

表 5.8　不同 WIP/产能水平下的 Move/产能与 CT 的敏感度分析表

Slope	$\frac{WIP}{产能}$	$\Delta\left(\frac{WIP}{产能}\right)$	$\frac{Move}{产能}$	$\Delta\left(\frac{Move}{产能}\right)$	CT	Δ(CT)
1.514	1.500		5.259		2.290	
1.696	1.577	0.077	5.383	0.124	2.295	0.005
1.804	1.654	0.077	5.518	0.135	2.309	0.014
1.838	1.731	0.077	5.659	0.141	2.331	0.022
1.798	1.808	0.077	5.799	0.140	2.362	0.031
1.683	1.885	0.077	5.934	0.135	2.402	0.040
1.495	1.962	0.077	6.057	0.123	2.451	0.049
1.232	2.039	0.077	6.162	0.105	2.509	0.058
0.896	2.116	0.077	6.245	0.082	2.575	0.066
0.485	2.193	0.077	6.298	0.054	2.651	0.075
0	2.270	0.077	6.317	0.019	2.734	0.084

通过本案例所提出的 CT 或 Move 的预测与控制的分析架构,可经由 SOM 拓扑图与属性间的散布图形来了解数据中所隐藏的生产形态与规则。在将预测模式应用于生产线之前,需经过敏感度分析过程来检定所构建模式的稳健性。此外,亦可借由不同水平的 WIP 所产生的 CT 预测时间或 Move 预测数目来观察斜率的变动情形。

5.7.5　案例小结

本案例提出的方法能较准确地推导出该厂的生产能力表现曲线;即使该厂的生产力有

剧烈变动时，本预测模型仍可有效控制预测误差，同时在数日内重新校正。因此，可借此预测周期时间的数据挖掘架构控管周期时间与产出，以提供管理者作为产能计划与需求管理的最佳决策基础。

制造管理数据受到较多人为管理因素的干扰，本研究发展多项式回归模型与 BPNN 模型，以建立高准确率的周期时间预测模型，一方面利用多项式回归，以建立目标函数和主要输入变量的因果关系，并提供合理的解释，以处理其中主要的变化趋势；另一方面，对于多项式回归无法完全解释的残差和变异，则借助人工神经网络高预测力的优点，以提升模型整体预测能力，并以反馈的方式，利用数学规划模式找出最适的 WIP 水平，以提供产能规划相关决策的评估依据。

在实际应用中，个案公司的产能及其他生产条件在过程中可能会有产能扩充或设备转换率等重大变化，以至于影响到实证构建模式的稳健性，因此数据的搜集及分析应持续进行，借由不断地数据挖掘工作，提供管理阶层实时有效的决策支持，以提升半导体厂制造管理与系统的整体产出绩效。

5.8　结论

当问题过于复杂、难以用数学模式计算，不需特别假设的人工神经网络就变得非常有用，借由学习的过程处理复杂的问题，许多种不同类型的网络形态也因应不同的问题类型而产生，如信号分类、语音识别转换、药物应用、债务分析与信用卡使用及投资贸易等。此外，人工神经网络具有高度的学习能力，对于高维度或非线性等复杂不易建立明确的数学关系模型的问题具有较佳的预测能力，即使在有少量噪声数据下，仍可有效运作。

人工神经网络模型需要谨慎的应用，传统人工神经网络不具备自动筛选变量的能力，当预测变量过多时可能造成网络结构过大，但实际上并非所有预测变量均对反应变量具有显著影响，用户可考虑结合决策树分析、统计检定方法，或其他维度缩减的方法先筛检变量，以降低数据维度。

如何决定最佳的人工神经网络参数，以避免得到局部最佳解而非全局最佳解亦为重要的议题，即使可采用不同的参数设定，例如学习率或惯性因子以试着得到近似的最佳解，然而也很难保证得到的结果一定是全局最佳解。此外，当问题具有大量的预测变量个数时，相较于其他分类或分群算法，人工神经网络需要较长的计算时间建立模型，也可能造成因模型重建造成的延迟。因此，如何克服人工神经网络模型的解释能力与计算时间上的落差，是实际应用人工神经网络方法时需要面对的挑战。

问题与讨论

1. 试问人工神经网络有何优缺点？可应用的问题类型有哪些？
2. 请解释构成人工神经模型的基本元素？
3. 人工神经网络在建立训练模型时，为什么需对输入与输出数据进行归一化？
4. 人工神经网络模型需要设定的项目包括隐藏层数目、隐藏藏单元个数、学习次数、学习率等参数，试举出一种以上决定参数的方法？

5. 承上题,各参数设定值的不同是否会影响训练模型的结果?

6. 假设一分类问题,输入变量为 I_1 与 I_2,输出变量为 O,对应的数据如下表:

I_1	I_2	O
−1	1	1
0	0	0
1	−1	0
1	0	1
0	1	1
1	1	0

(1) 请画出上表中的网络图,并给定相关的初始参数值。

(2) 请利用反向传播人工神经网络说明一次学习的过程。

7. 请举一实际范例说明自组织映射网络图的应用。

8. 假设欲将 4 个输入向量(1,1,0,0)、(0,0,1,1)、(0,0,1,0)、(0,1,0,0),利用 SOM 算法进行分群,分群个数为两群,请说明此过程如何进行。

9. 下表为 10 个 2×3 的输入图像为例,假设警戒门槛值 $\rho=0.5$,

(1) 利用 ART1 计算其图样的分群结果。

(2) 如果 $\rho=0.25$,对 ART1 的影响为何?请比较不同警戒门槛值下的 ART1 结果。

10 种输入图像

1	2	3	4	5
■■■ □□□	□□□ ■■■	■□□ □■■	□□■ ■■□	■□■ □■□
6	7	8	9	10
□■□ ■□■	□■□ ■■■	■□□ ■■■	□□■ ■■■	□□□ □■□

10. ART 人工神经网络要如何兼顾稳定性与可塑性?

11. 应用 ART 网络时若发现得到的聚类数目过多,可能的原因为何?有何可能的解决办法?

第 6 章

聚 类 分 析

6.1　聚类分析法简介

聚类分析(**clustering analysis**)是依据数据相似度或相异度而将数据分群归属到数个聚类(clusters)的方法;使得同一群内的数据或个体相似程度大,而各群之间的相似程度小。同一组样本有时会因为不同的目的、数据输入方式、所选的分群特征或数据属性,形成不同的分群结果。例如,图 6.1(a)的数据,可以根据某些特征和准则,将数据分成 3 个(图 6.1(b))或 4 个(图 6.1(c))聚类。另一方面,**分类**(**classification**)则是根据已知或所给定目标数据的类别,找出其分类属性,建立分类规则或模式,将数据分类至所对应的目标类别。

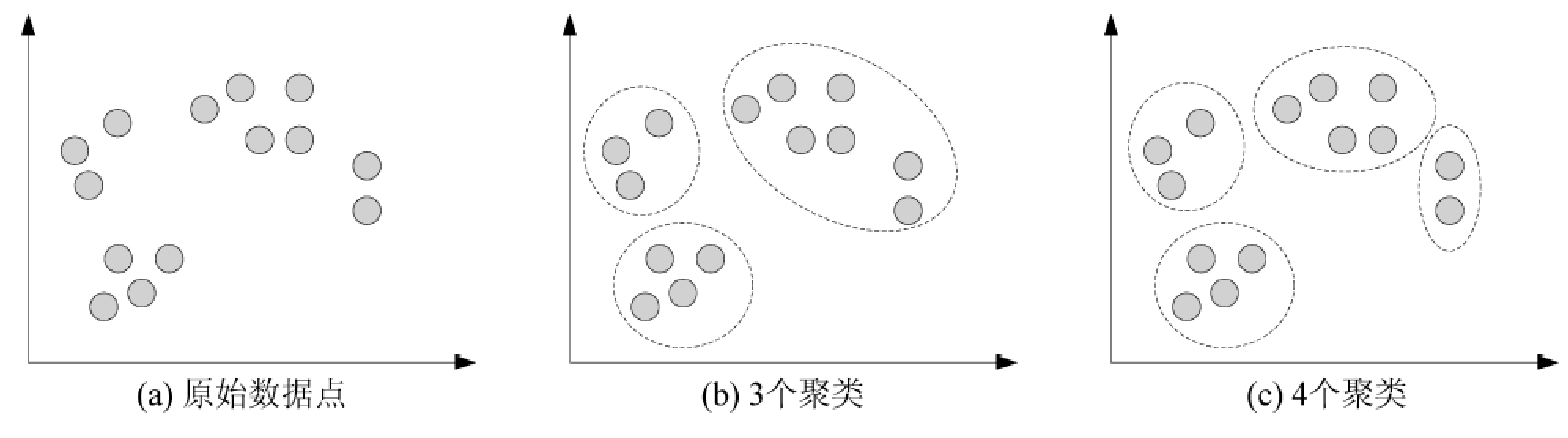


图 6.1　不同的分群结果

聚类分析是分群以找出各子聚类数据背后可能隐藏的特征、样型或关联现象。聚类分析事先并不知道聚类数目,而分群结果的特征及其所代表的意义仅能事后加以解释。因此,聚类分析可视为无监督式学习;而分类方法则视为监督式学习。

聚类分析应用的领域相当广泛。例如,根据顾客基本数据和事务数据将顾客分群,定义并分析不同类型顾客的消费行为模式,以设计定制化的营销方案;或是通过聚类分析将信用卡使用行为分为不同群组样型,以分析信用卡异常消费的情形,避免盗刷所造成的损失。在制造业,可依据机台的特征、功能等的相似程度,将机台分为可以相互替代和备援(backup)的聚类,以提升作业效率并维持良率(Chien & Hsu,2006)。在网络营销中,可将性质或特性相仿的网页予以分类,增快网页搜索速度,并根据浏览行为和客户聚类分析作客户消费行为预测和搭配营销。

此外,聚类分析也常常与其他算法整合,将分群结果输入后续的分析中。例如,提取各聚类的特征作为后续分类的准则;或在数据准备时,运用聚类分析决定群组并将离散数据以代码表示。

6.1.1 聚类分析的阶段

聚类分析主要包括以下四个阶段：

（1）数据准备与分群特征选取：根据问题特性、数据类型及所选择的分群算法等，自搜集的变量中选取具代表性的变量作为分群特征属性。

（2）相似度计算：选择衡量相似度的方式，如距离、相关系数等。在选择衡量相似度的方式时，需考虑数据的类型以及后续使用的分群算法，例如，在类别尺度中，选用欧式距离可能会造成数据尺度的误用。

（3）分群算法：为整个聚类分析中最重要的阶段，主要为利用分群算法将数据分组，有些分群算法可能需要自行决定群数，例如，划分聚类分析算法可由用户自行决定或利用其他方式决定适当的分群个数。

（4）分群结果评估与解释：当分群结束后需检查分群结果是否合理。例如，聚类间的距离是否过大、该数据是否适用所选用的分群算法，若发现有不合理的地方，则需重新审视前三个阶段是否有问题。另外，由于分群后的结果可能作为另一个方法的输入数据，因此可能需要对聚类结果进行定义或命名。

本章主要介绍如何衡量数据间或聚类间的相似度、分群算法的种类，而具体的步骤则依不同领域可能会因为输入的数据及所选择的分群算法的差异而有所不同。有关数据搜集、数据处理与特征选取可参阅第 2 章的详细介绍，而结果的评估与解释往往需要与领域专家进一步讨论，以检验分析模式的效度。

6.1.2 相似度的衡量

相似度（similarity）代表对象或个体间的近似或相关程度，可作为决定分群的依据，以及个体在不同聚类间的归属。相似度的数值越大，表示数据间关联的程度越高，应归类于同一聚类；反之，若相似度的数值越小，表示数据间关联的程度越低，则应归类于不同聚类。

假设有 N 笔数据，每笔数据有 P 个变量，而 x_{ij} 表示第 i 笔数据在第 j 个变量的值，以下将介绍多个变量下如何衡量数据相似度的方法。

1. 距离

"距离"常用来衡量两笔数据或两个体在一维或多维变数下的相异程度。距离越大，表示相异度越大，反之则越小。常用的距离衡量方式如下。

（1）欧氏距离（Euclidean distance）

欧式距离为常用的距离衡量方式，如式（6.1），表示多维空间下两个数据点间的几何距离，如图 6.2 中的虚线。

$$D_{(y_1,y_2)} = \sqrt{\sum_{j=1}^{P}(x_{1j}-x_{2j})^2} \tag{6.1}$$

其中，$D_{(y_1,y_2)}$ 表示两个数据点间的欧式距离。

然而，实际使用欧式距离衡量数据点之间的差异程度时，因为开根号计算较为不易，分析时亦可改用欧式距离的平方（squared Euclidean distance）代替。

（2）**曼哈顿距离（Manhattan distance）**

曼哈顿距离是另一个常用来测量距离的方式，又称为城市街道距离（city-block distance），定义为各变量差距的绝对值之和，如图 6.2 中的实线，衡量公式如式（6.2）：

$$D_{(y_1,y_2)} = \sum_{j=1}^{P} |x_{1j} - x_{2j}| \tag{6.2}$$

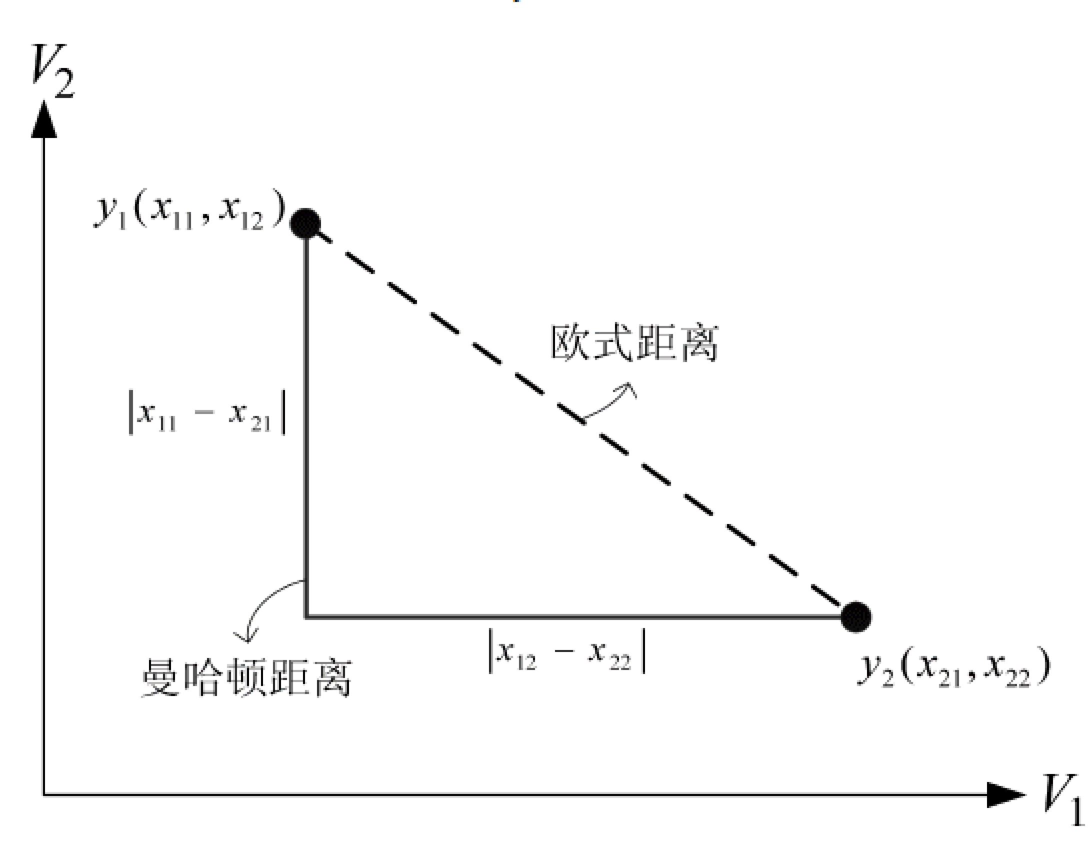


图 6.2　欧式距离与曼哈顿距离示意图

（3）**闵氏距离（Minkowski distance）**

闵氏距离可视为欧式距离与曼哈顿距离的通式，当闵氏距离的参数 $n=1$ 时即为曼哈顿距离；当 $n=2$ 时，即为欧式距离。其中，n 为正整数，如式（6.3）：

$$D_{(y_1,y_2)} = \left(\sum_{j=1}^{P} |x_{1j} - x_{2j}|^n\right)^{1/n} \tag{6.3}$$

（4）**加权距离（weighted distance）**

当各变量的重要性不同时，可给定相对权重 w_j，以衡量加权距离。以欧式距离为例说明，如式（6.4）：

$$D_{(y_1,y_2)} = \sqrt{\sum_{j=1}^{P} w_j\,(x_{1j} - x_{2j})^2} \tag{6.4}$$

其中，所有加权权重 w_i 总和为 1；当权重都相同时，加权距离就等价于欧式距离。

（5）**标准化距离（normalized distance）**

在衡量距离时，若不同维度数据的衡量尺度或单位不同时，衡量结果变异较大的变量可能会凌越（dominate）最后的结果。举例来说，若以年资（单位：年）与薪水（单位：元）作为衡量两人之间距离的特征变量，由于薪水的变异较大，因此薪水的差异会决定最后的距离。要解决数据在不同尺度上的差异，可先对变量进行标准化，根据平均数与标准差将数据转换至同一比较基准（详细方法可参阅第 2 章），即可避免变量间因尺度不同而导致数据分布范围差异过大的问题。标准化的优点是，转换后的数据可用以检测异常值。

（6）**马氏距离（Mahalanobis distance）**

若所欲衡量的变量间除了尺度差异，变量间也具有相关性时，可改用马氏距离公式以衡量数据点之间的距离，如式（6.5）：

$$D_{(y_1,y_2)} = (\boldsymbol{x}_1 - \boldsymbol{x}_2)'\,\boldsymbol{S}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2) \tag{6.5}$$

$D_{(y_1,y_2)}$表示群体间的马氏距离,$\boldsymbol{x}_1=(x_{11},\cdots,x_{1P})^{\mathrm{T}}$ 与 $\boldsymbol{x}_2=(x_{21},\cdots,x_{2P})^{\mathrm{T}}$ 均为 $P\times1$ 的向量,$\boldsymbol{S}$ 为 P 个变数的共变异矩阵。当变量间没有相关性(相关系数等于 0),并且所有变量的方差都为 1 时,马氏距离即等于标准化的欧式距离。马式距离的计算虽然较为繁复,但其优点是可考虑变数间的相关性。

2. 相关系数

(1) 皮尔逊相关系数

相关系数(correlation coefficient)衡量两随机变量的变动方向与程度大小以描述其相关性,也可作为两变量的相似度量测。在连续型数据中最常使用的是皮尔逊相关系数(Pearson correlation coefficient),又称线性相关系数。对 V_1、V_2 两变量,假设 N 组数据 $(x_{11},x_{12}),(x_{21},x_{22}),\cdots,(x_{N1},x_{N2})$,则其相关系数 $\boldsymbol{r}_{(V_1,V_2)}$ 的定义如式(6.6):

$$\boldsymbol{r}_{(V_1,V_2)}=\frac{\sum_{i=1}^{N}(x_{i1}-\bar{x}_{\cdot 1})(x_{i2}-\bar{x}_{\cdot 2})}{\sqrt{\sum_{i=1}^{N}(x_{i1}-\bar{x}_{\cdot 1})^2}\sqrt{\sum_{i=1}^{N}(x_{i2}-\bar{x}_{\cdot 2})^2}} \tag{6.6}$$

由式(6.6)可知,相关系数与单位无关;且相关系数介于-1 到$+1$ 之间。当 $\boldsymbol{r}_{(V_1,V_2)}>0$ 表示 V_1 增加时,V_2 也增加;$\boldsymbol{r}_{(V_1,V_2)}<0$ 表示 V_1 增加时,V_2 则减少。一般而言,$0\leqslant|\boldsymbol{r}_{(V_1,V_2)}|\leqslant0.3$ 表示两变量为低相关性,$0.3<|\boldsymbol{r}_{(V_1,V_2)}|\leqslant0.7$ 表示两变量为中相关性,$0.7<|\boldsymbol{r}_{(V_1,V_2)}|\leqslant1$ 表示两变量为高相关性。

(2) 等级相关系数

针对顺序尺度数据则可用斯皮尔曼等级相关系数(Spearman's rank correlation coefficient)r_S,如式(6.7):

$$r_S=1-\frac{6\sum_{i=1}^{N}[R(x_{i1})-R(x_{i2})]^2}{N(N^2-1)} \tag{6.7}$$

其中,$R(x_{i1})$与 $R(x_{i2})$代表 V_1、V_2 两变量第 i 笔数据的顺序,r_S 越大代表两变量样本数据间的顺序一致性越高,并非其样本数据值具有高度相关。若所有成对数据的顺序均相同,则 $r_S=1$,代表两变量等级数据具有高度一致性。

3. 二元关联系数

当类别变量仅有两个状态:无或有(0 或 1),称为二元变量或布尔变量。例如,晶圆在某道制程中是否有经过该机器,其中,0 表示没有,1 表示有。对两个二元变量形态的数据进行聚类分析时,假设各变量的重要性相同,以表 6.1 的列联表(contingency table)为例。

表 6.1 2×2 列联表

V_1	V_2		
	0	1	加总
0	r	s	$r+s$
1	t	u	$t+u$
加总	$r+t$	$s+u$	N

其中，r 表示变量 $V_1=0$ 且变数 $V_2=0$ 的数据笔数，s 表示变量 $V_1=0$ 且变数 $V_2=1$ 的笔数，t 表示变量 $V_1=1$ 且变数 $V_2=0$ 的笔数，u 表示变量 $V_1=1$ 且变数 $V_2=1$ 的笔数。N 为总数据笔数。衡量类别数据的相似度可以用简单比对系数（simple matching coefficient，SMC），如式（6.8）：

$$S(V_1,V_2)=\frac{r+u}{r+s+t+u} \tag{6.8}$$

若两个变量的重要性有所不同时，则表示该二元变量为不对称的（asymmetric）。例如，1 表示一片晶圆经过 A 制程，0 表示该晶圆没有经过 A 制程，但实际上对于工程师而言，经过该制程的数据较没有经过该制程更具有意义，若用简单比对系数则可能无法表现其中的差异，因此，变量 $x=0$ 且变数 $y=0$ 配对的次数是不被考虑的，可以用 **Jaccard 系数**如式（6.9）来衡量相似度：

$$J=\frac{u}{s+t+u} \tag{6.9}$$

6.1.3　聚类分析方法

常用的聚类分析算法，包括层次聚类分析、划分聚类分析、以密度为基础和以模式为基础的聚类方法等，说明如下。

1. 层次聚类分析

层次聚类分析（hierarchical clustering）是对数据点进行层次的聚类，而用树形图（dendrogram）表示各聚类中所包括的数据点，树形图的根节点仅包含单一聚类，代表所有数据点均落在同一聚类中，而树形图中的叶节点皆各自为单一聚类，代表各数据点均为独立聚类。

层次聚类分群方式可分为**凝聚（agglomerative）**与**分裂（divisive）**两种。凝聚的方法是由下而上（bottom-up），先将各样本点视为单独的聚类，在接下来的每一步骤将最相似的聚类合并，直到所有的数据点均合并到同一聚类中或达到所规定的停止条件为止，大部分的层次聚类算法均属于这一类；分裂的方法是一种由上而下（top-down）的方法，一开始先将所有个体凝聚为一个大聚类，之后的每一步骤，从原有的聚类中挑选一个聚类，依据相异度的差别再分裂为两个较小的聚类，直到每个数据点各自成为一个独立的聚类或达到所规定的停止条件为止。一般而言，凝聚方法较分裂方法更常使用（Kantardzic，2003）。

2. 划分聚类分析

划分（partition）是先选择数个不同的起始聚类中心点，每一个数据点只会被分到一个聚类，首先所有样本数据均计算与每个中心点的距离或相似度，而每个样本会根据具有最小距离或相似度的结果将其划分至该聚类，往往以平方误差（squared error）为衡量划分结果，具有最小平方误差的划分即为最终的分群。

3. 以密度为基础的方法

层次聚类分析与划分聚类分析大多以数据点或聚类间的距离作为分群依据，然而，这样的衡量尺度只能得到球状的分群结果。

若数据点的分布为任意形状，则应考虑到所获得数据的紧密程度，改用基于密度的聚类

分析法，以得到任意形状的聚类。

4. 以模式为基础的方法

以模式为基础的方法是将数据根据模型予以配适而产生聚类，例如，以第 5 章的自组织映射图网络为基础，将数据点投射至二维平面来进行聚类分析。

6.2 层次聚类分析法

层次聚类分析法的每一个新聚类均是由下一阶层的聚类所凝聚或上一阶层的聚类分裂而得，其形成的方式就像一个树状结构。凝聚式层次分群算法是将所有数据视为单一聚类，并计算所有聚类内的距离矩阵，再将最近的两笔数据合成一群，重新计算聚类间的相似度，直到所有数据都在一个聚类内为止。

层次聚类算法是以两聚类间的**相近程度**（**proximity**）为基础，根据不同距离的选用，表示两聚类的相似程度。

几个常用来衡量聚类间的相近程度公式，说明如下：

最小距离（minimum distance）：$$D_{\min}(C_i,C_j)=\min_{a\in C_i,b\in C_j} D_{(a,b)} \tag{6.10}$$

最大距离（maximum distance）：$$D_{\max}(C_i,C_j)=\max_{a\in C_i,b\in C_j} D_{(a,b)} \tag{6.11}$$

平均距离（average distance）：$$D_{\text{average}}(C_i,C_j)=\frac{1}{n_i n_j}\sum_{a\in C_i}\sum_{b\in C_j} D_{(a,b)} \tag{6.12}$$

中心值距离（centroid distance）：$$D_{\text{centroid}}(C_i,C_j)=D_{(m_i,m_j)} \tag{6.13}$$

其中，m_i 与 m_j 分别表示聚类 C_i 与 C_j 的中心值，n_i 与 n_j 分别表示聚类 C_i 与 C_j 的数据点个数，$D_{(a,b)}$ 表示两样本点间的距离，可以使用的距离衡量方式有欧式距离或曼哈顿距离等，更进一步的说明可见图 6.3。

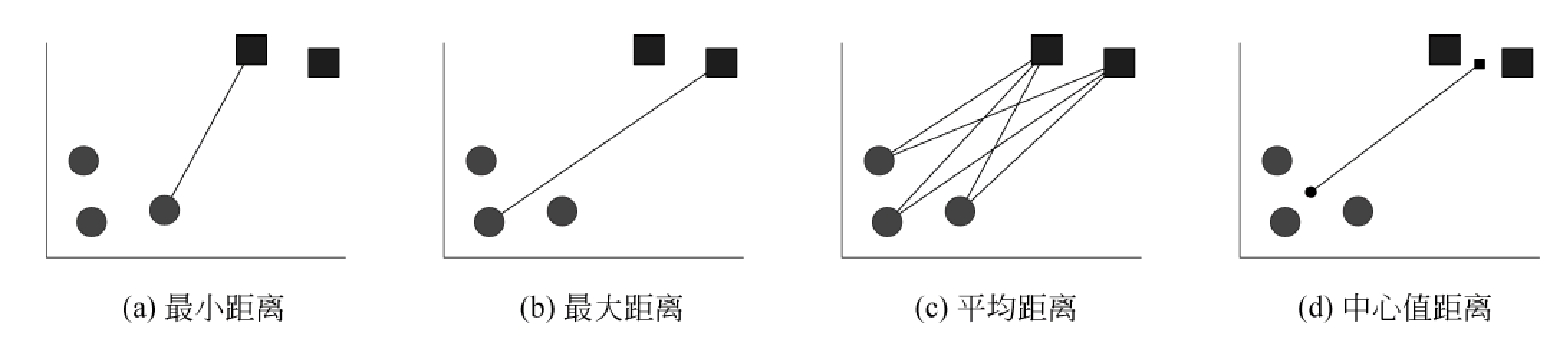


图 6.3 距离示意图

［**范例 6.1**］ 为 7 笔观察值的 V_1 与 V_2 数据（如表 6.2），为方便计算，以欧式距离平方作为衡量相似度的依据，可计算出各数据点间的欧式距离平方如表 6.3 所列。假设现在有三个聚类，分别是聚类 $A=\{1,3,6\}$，聚类 $B=\{2,4\}$，聚类 $C=\{5,7\}$，聚类 A 与 B 间共有 6 个距离，分别为：$D_{1\&2}=233$、$D_{1\&4}=261$、$D_{3\&2}=149$、$D_{3\&4}=169$、$D_{6\&2}=80$、$D_{6\&4}=104$。

表 6.2 ［范例 6.1］观察值

观察值	V_1	V_2
y_1	14	15
y_2	22	28

续表

观察值	V_1	V_2
y_3	15	18
y_4	20	30
y_5	30	35
y_6	18	20
y_7	32	30

表 6.3　[范例 6.1]欧式距离平方

序号	1	2	3	4	5	6	7
1	0	233	10	261	656	41	549
2	233	0	149	8	113	80	104
3	10	149	0	169	514	13	433
4	261	8	169	0	125	104	144
5	656	113	514	125	0	369	29
6	41	80	13	104	369	0	296
7	549	104	433	144	29	296	0

若使用最小距离作为聚类间相近程度的衡量，则两聚类间的距离为 $D_{\min(C_A,C_B)}=D_{6\&2}=80$。

若使用最大距离作为聚类间相近程度的衡量，聚类 A 与聚类 B 间的距离为 $D_{\max(C_A,C_B)}=D_{1\&4}=261$。

若使用平均距离作为聚类间相近程度的衡量，则聚类 A 与聚类 B 的距离为 $D_{\text{average}(C_A,C_B)}=\dfrac{D_{1\&2}+D_{1\&4}+D_{3\&2}+D_{3\&4}+D_{6\&2}+D_{6\&4}}{6}=166$。

若使用中心值距离作为聚类间相近程度的衡量，聚类 A 的中心为 $\left(\dfrac{14+15+18}{3},\dfrac{15+18+20}{3}\right)=\left(\dfrac{47}{3},\dfrac{53}{3}\right)$，聚类 B 的中心为 $\left(\dfrac{22+20}{2},\dfrac{28+30}{2}\right)=(21,29)$，则聚类 A 与聚类 B 的欧式距离为 $D_{\text{centroid}(C_A,C_B)}=\left(21-\dfrac{47}{3}\right)^2+\left(29-\dfrac{53}{3}\right)^2=156.89$。

常见的层次聚类分析方法包括：**单一连结法(single linkage method)**，以两聚类间数据点中的最小距离来表示两聚类的距离及两群数据的邻近程度；**完全连结法(complete linkage method)**，以两聚类间数据点的最大距离来表示两聚类的距离及两群数据的邻近程度；**平均连结法(average linkage method)**，衡量聚类内所有点到另一个聚类内所有点的距离平均来表示两聚类的邻近程度，以避免聚类之间的距离衡量受噪声影响；**中心点连结法(centroid linkage method)**，以两聚类的中心点距离作为衡量两聚类的距离，以表示其邻近程度。

以[范例 6.1]为例，利用单一连结法说明层次聚类分析的计算，起初所有数据皆属于单一聚类，而数据点 2 与数据点 4 最接近，所以将两点合并为一聚类，重新计算各聚类间数据点的最小距离如表 6.4 所示。而数据点 1 与数据点 3 最为接近，因此将两点合并为新的聚

类，迭代，直到将所有数据点均合并至同一聚类中为止。

表 6.4 单一连结法：合并 2 和 4 后的欧式距离

序号	1	2&4	3	5	6	7
1	0	233	10	656	41	549
2&4	233	0	149	113	80	104
3	10	149	0	514	13	433
5	656	113	514	0	369	29
6	41	80	13	369	0	296
7	549	104	433	29	296	0


最后，聚类 AB 与聚类 C 在距离为 104 时合并为一群，如图 6.4 所示。

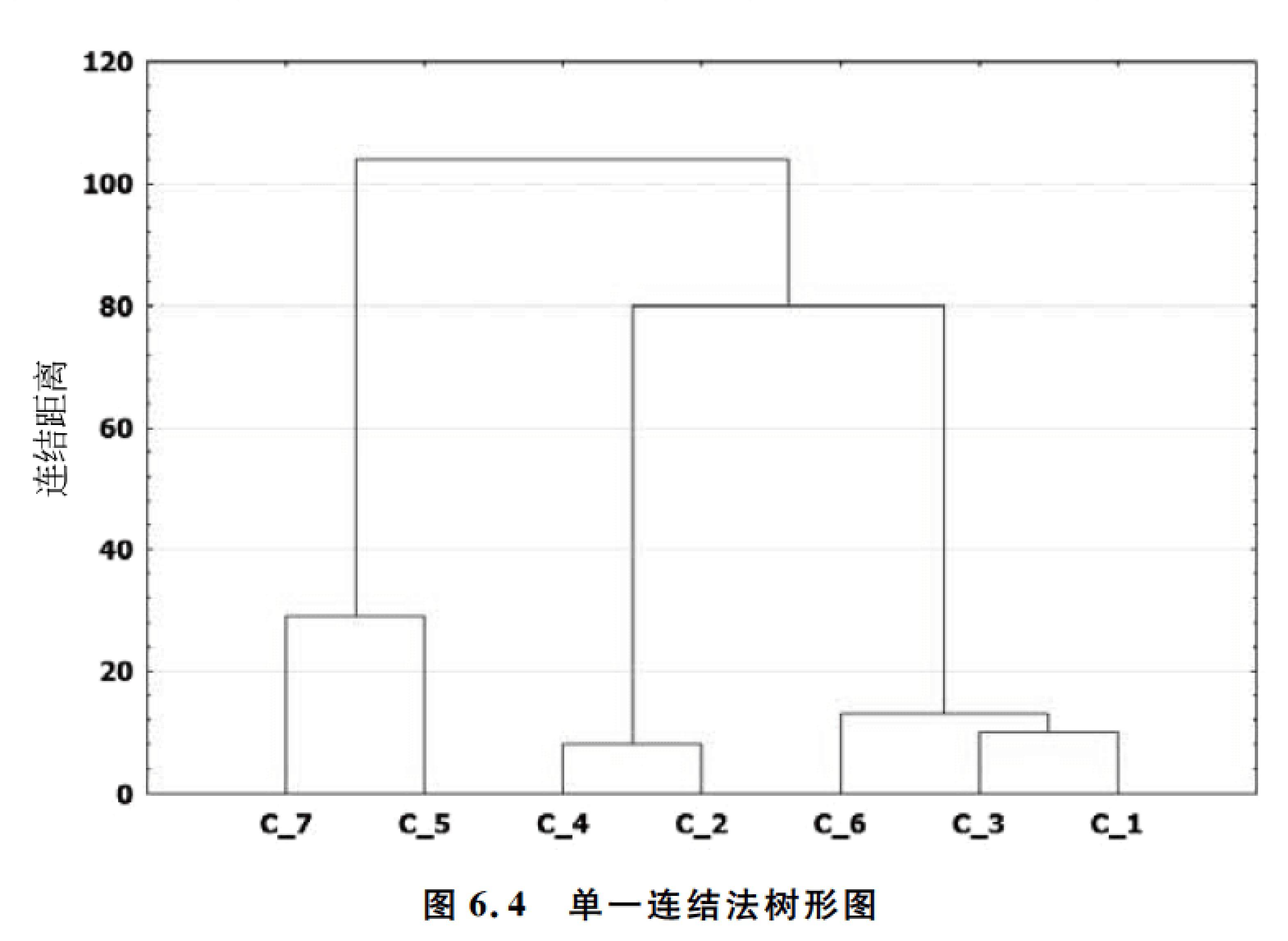


图 6.4 单一连结法树形图

另一种层次聚类分析方法**沃德法(Ward's method)**，以衡量各聚类间组内变异作为衡量聚类相似度的分群方法(Ward，1963)，依序将所有聚类合并，反复计算与合并每一阶段中最小聚类的组内变异，直到所有数据均合并为一群为止，使聚类内数据的同构型(homogeneity)最大化，亦即聚类内变异最小化，衡量的方法以和方差(sum of squared errors，SSE)如式(6.14)：

$$\mathrm{SSE} = \sum \sum (x_{ij} - \bar{x}_{\cdot j})^2 \tag{6.14}$$

以[范例 6.1]为例，起始时所有数据皆属于单一聚类，因此组内变异和为 0。在步骤二中发现数据点 2 与数据点 4 合并后的组内变异和最小，所以将两点合并为一新聚类，再重新计算各聚类间组内的变异如表 6.5 所示，从中可发现数据点 1 与数据点 3 合并后所增加的组内变异最少，所以再将两点合并为新的聚类，如此迭代，直到将所有数据点均合并至同一聚类中为止，分群结果如图 6.5。

表 6.5　沃德法计算次数 2

序号	聚类组合						组内变异和
	C_1	C_2	C_3	C_4	C_5	C_6	
1	1&2	3	4	5	6	7	116.5
2	1&3	2	4	5	6	7	5
3	1&4	2	3	5	6	7	130.5
4	1&5	2	3	4	6	7	328
5	1&6	2	3	4	5	7	20.5
6	1&7	2	3	4	5	6	274.5
7	2&3	1	4	5	6	7	74.5
8	2&4	1	3	5	6	7	4
9	2&5	1	3	4	6	7	56.5
10	2&6	1	3	4	5	7	40
11	2&7	1	3	4	5	6	52
12	3&4	1	2	5	6	7	84.5
13	3&5	1	2	4	6	7	257
14	3&6	1	2	4	5	7	6.5
15	3&7	1	2	4	5	6	216.5
16	4&5	1	2	3	6	7	62.5
17	4&6	1	2	3	5	7	52
18	4&7	1	2	3	5	6	72
19	5&6	1	2	3	4	7	184.5
20	5&7	1	2	3	4	6	14.5
21	6&7	1	2	3	4	5	148

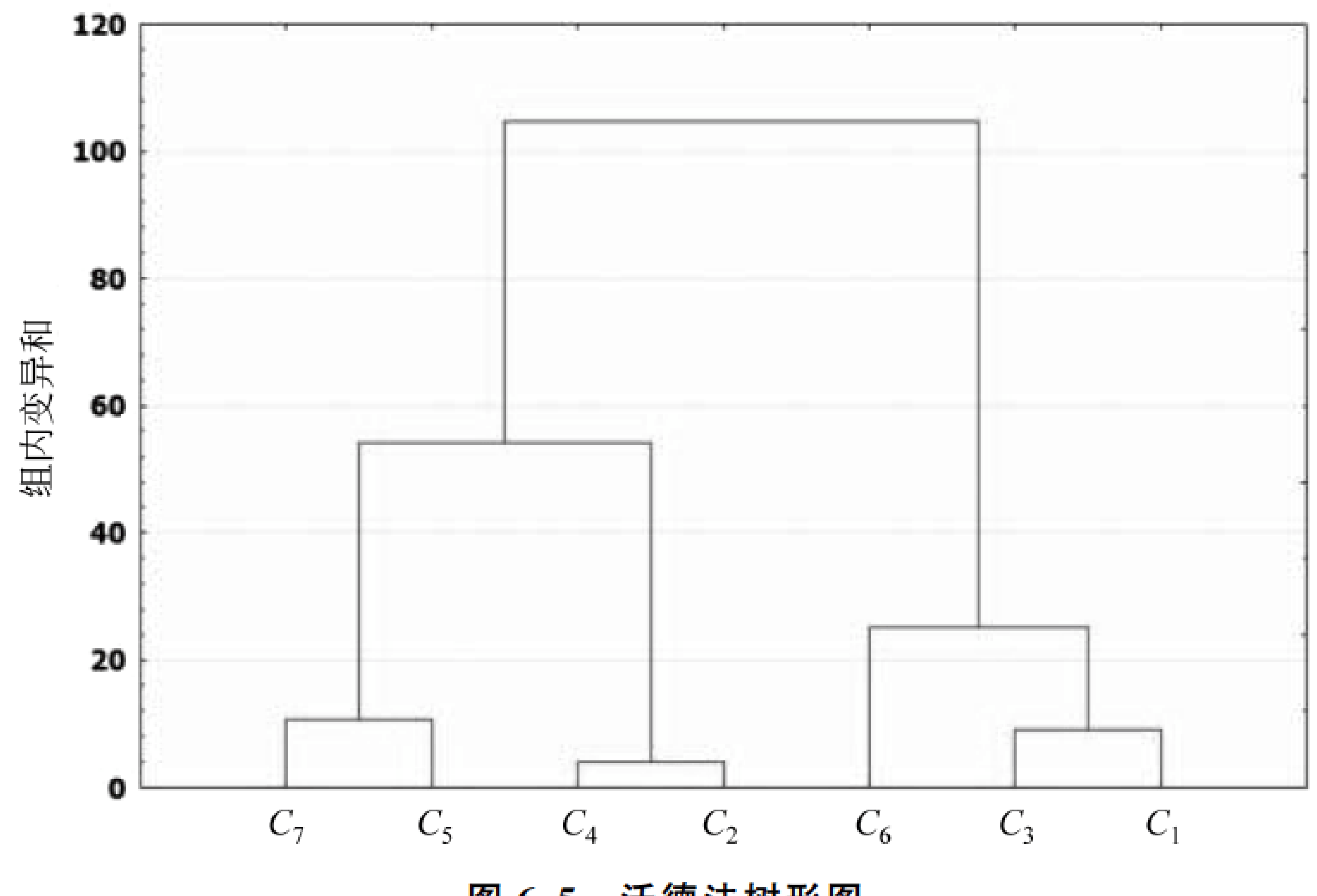


图 6.5　沃德法树形图

6.3　划分聚类分析法

划分聚类方析法将个体分成 k 组划分区域，每个区域代表满足特定条件或特征下的群组，假设有 N 笔数据在依据 P 种特征下，要被分到 k 个聚类中$\{C_1,C_2,\cdots,C_k\}$，每一聚类 C_l 包含 n_l 笔数据，且每一笔数据只被分到其中一个聚类，即 $\sum_{l=1}^{k} n_l = N$，其中，$l=1,2,\cdots,k$，假设$\boldsymbol{x}_{il}$为在聚类 C_l 中的第 i 笔数据向量，$\boldsymbol{m}_l$ 代表聚类 C_l 数据的中心向量，一般可采用聚类的平均值或聚类的中心值。根据聚类 C_l 内数据点与聚类中心$\boldsymbol{m}_l$ 的距离差异平方可计算如式(6.15)：

$$e_l^2 = \sum_{i=1}^{n_l} (\boldsymbol{x}_{il} - \boldsymbol{m}_l)^{\mathrm{T}} (\boldsymbol{x}_{il} - \boldsymbol{m}_l) \tag{6.15}$$

再考虑所有 k 个聚类，可定义所有 k 个聚类内的总方差如式(6.16)：

$$E = \sum_{l=1}^{k} e_l^2 \tag{6.16}$$

因此，根据所定义的总平方误差 E，具有最小总聚类变异之划分 k 个聚类的结果为建议的分群。以下介绍 K 平均法与 K 中心点法等划分聚类分析。

6.3.1　K平均法

K* 平均法**(K*-means method**)是将数据分割成 K 个互不相交的聚类，当数据点与该聚类中心的相似度高于其他聚类时，则归类于该聚类中，若与其他聚类中心的相似度相较之下高于原有聚类中心时，则将该数据点归属于新聚类，而再以新聚类中所计算出的新的平均值为中心，如此反复计算直到切割子集的结果收敛为止。K 平均法的目标为使每个数据点到所属聚类中心的总距离变异平方和最小，在规定聚类中心时则是以该聚类中数据点的平均值作为该聚类的中心，如式(6.17)所示：

$$E = \sum_{l=1}^{K} \sum_{i=1}^{n_k} (\boldsymbol{x}_{il} - \boldsymbol{m}_l)^{\mathrm{T}} (\boldsymbol{x}_{il} - \boldsymbol{m}_l) \tag{6.17}$$

其中，$\boldsymbol{x}_{il}$代表聚类 C_l 中的某一笔数据，$\boldsymbol{m}_l$ 代表聚类 C_l 的平均值，E 为总距离变异平方和。

K 平均法的步骤如下：

(1) 首先随机选取 K 笔数据点作为 K 个起始聚类中心值。

(2) 将剩下的每一笔数据分配到离聚类中心最近的聚类中，并根据聚类中的数据点，重新计算各聚类的平均值。

(3) 计算数据点到聚类中心的距离，若发现总距离变异平方和下降，则表示聚类中心有所改变，需将数据点重新分配到新的聚类。

(4) 这过程会不断持续，直到总距离变异不再下降或达到所设定的计算次数为止。

以[范例 6.1]为例说明 K 平均法的步骤。设聚类数 $K=3$，以欧式距离平方作为衡量相似度的依据，先随机选取数据 1、4、6 作为起始聚类中心，如表 6.6 与图 6.6(a)，接着计算各数据点至各聚类中心的距离，将数据点分配至距离最近的聚类中心(如表 6.7)，再根据该聚类中的数据点，重新计算各聚类中心(如表 6.8)，并再次分群，如图 6.6(b)与表 6.9 所列，各

数据仍被归纳至原有聚类，表示聚类中心不变，所以停止继续分群。

表 6.6　起始聚类中心

聚类	V_1	V_2
A	14	15
B	20	30
C	18	20

表 6.7　*K* 平均法分群过程(初始重新分配)

序号	与聚类中心的距离			最小距离	分配的聚类
	聚类 *A*	聚类 *B*	聚类 *C*		
1	0	261	41	0	*A*
2	233	8	80	8	*B*
3	10	169	13	10	*A*
4	261	0	104	0	*B*
5	656	125	369	125	*B*
6	41	104	0	0	*C*
7	549	144	296	144	*B*

表 6.8　聚类中心(第一次重新分配)

聚类	V_1	V_2
A	14.5	16.5
B	26	30.75
C	18	20

表 6.9　*K* 平均法分群过程(第一次重新分配)

序号	与聚类中心的距离			最小距离	分配的聚类
	聚类 *A*	聚类 *B*	聚类 *C*		
1	2.5	392.06	41	2.5	*A*
2	188.5	23.56	80	23.56	*B*
3	2.5	283.56	13	2.5	*A*
4	212.5	36.56	104	36.56	*B*
5	582.5	34.06	369	34.06	*B*
6	24.5	179.56	0	0	*C*
7	488.5	36.56	296	36.56	*B*

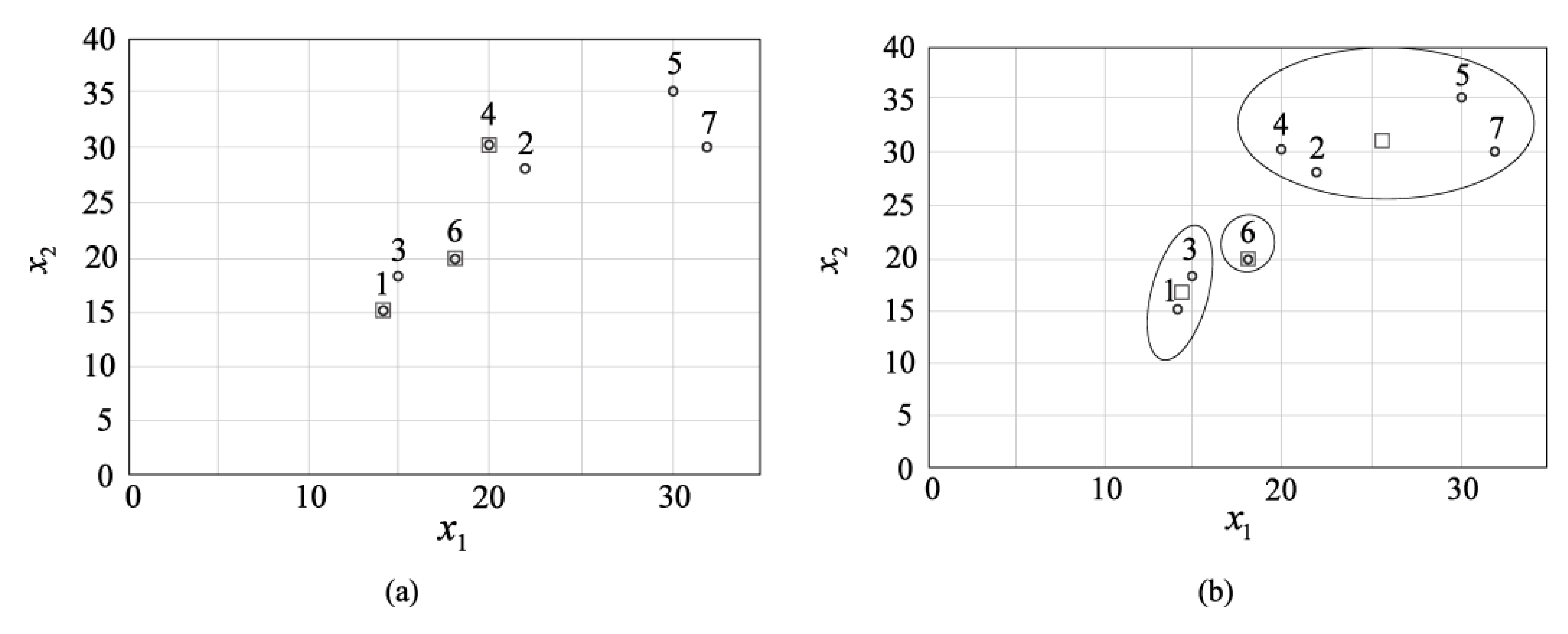


图 6.6　K 平均法对[范例 6.1]的分群过程

然而，若实际的分群结果为聚类 $A=(1,3,6)$，聚类 $B=(2,4)$，聚类 $C=(5,7)$，总距离变异平方和会较上面的结果小。由此可知，选择起始数据作为聚类中心可能会影响分群的结果。

K 平均法虽然已广泛使用在聚类分析上，但仍存在一些缺点：

- K 平均法无法直接处理类别型的数据（因无法求得数据的中心点），这类型数据可改用另一种划分聚类分析法 **K 众数法（K-mode）**进行分群。K 众数法是用简单配对相异度（simple matching dissimilarity）作为衡量相似度的指标，并以聚类的众数作为聚类的中心，用频率为基础（frequency-based）的方法来更新聚类的众数，详细内容可进一步参见（Huang，1998）。
- K 平均法必须事先决定聚类数目。聚类数目往往需由用户直接给定，或通过反复分析与验证，取得适当的群数。另外可利用两阶段的方式，也就是先用层次聚类分析算法决定聚类的数目，再利用 K 平均法重新将数据归类分群（Sharma，1996）。
- 分群结果容易受到离群值的影响。因为 K 平均法是以平均值作为聚类的中心，在计算时容易受到离群值的影响造成偏移，产生聚类分布上的误差。为了避免离群值影响分群结果，可改用 **K 中心点法**进行分群。
- 起始聚类中心选择的不同会造成不同的分群结果，若起始聚类中心的数据不够分散，可能会得到较差的聚类结果。
- 无法适用于所有的数据聚类形态，如 K 平均法无法处理非球状的聚类、数据大小差异很大的聚类，和数据密度不同的聚类。
- 当聚类间的特性非常相似时，在边界上的数据点只要有一点偏差，就可能从 A 聚类划分到 B 聚类。这类型的数据可改用柔性聚类（soft clustering）方法来处理。

6.3.2　K中心点法

K 中心点法（K-mediods method）与 K 平均法均使用距离作为衡量相似度的依据，并最小化数据点与聚类中心的总变异。然而，K 中心点法以聚类中最接近中心位置的数据点作为聚类的中心；K 平均法则使用全部数据的平均值作为聚类中心，因此，K 中心点法较不易受噪声与异常值的影响。

聚类变异衡量公式可修正如下：

$$E_k^2=\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ik}-\boldsymbol{x}_{mk})^{\mathrm{T}}(\boldsymbol{x}_{ik}-\boldsymbol{x}_{mk}) \tag{6.18}$$

其中，$\boldsymbol{x}_{ik}$代表聚类k中的某一数据点，$\boldsymbol{x}_{mk}$代表聚类k中最接近中心的数据点。

K中心点法的计算步骤与K平均法类似。围绕中心点划分法（partition around medoids，PAM）是其中较具代表性的方法之一（Kaufman & Rousseeuw，1990）。计算步骤如下：

（1）选取K个较具代表性的数据作为聚类的起始中心。

（2）依据距离的远近，将数据分配到最近的聚类。

（3）选取任一非聚类中心的数据点取代任一聚类中心，并计算总聚类距离改变量S。当$S<0$时，以该数据取代原有的聚类中心，而$S>0$时，则表示原有的聚类中心不需要被取代。

（4）重复步骤（3），直到确定所有数据点均无法取代任一聚类中心为止。以[范例6.1]为例，K中心点法的计算过程如表6.10至表6.12：

表6.10　K中心点法计算过程1（$K=3$）

序号	与聚类A(1)的相异度	与聚类B(4)的相异度	与聚类C(6)的相异度	最小相异度	分配的聚类
1	0	261	41	0	A
2	233	8	80	8	B
3	10	169	13	10	A
4	261	0	104	0	B
5	656	125	369	125	B
6	41	104	0	0	C
7	549	144	296	144	B

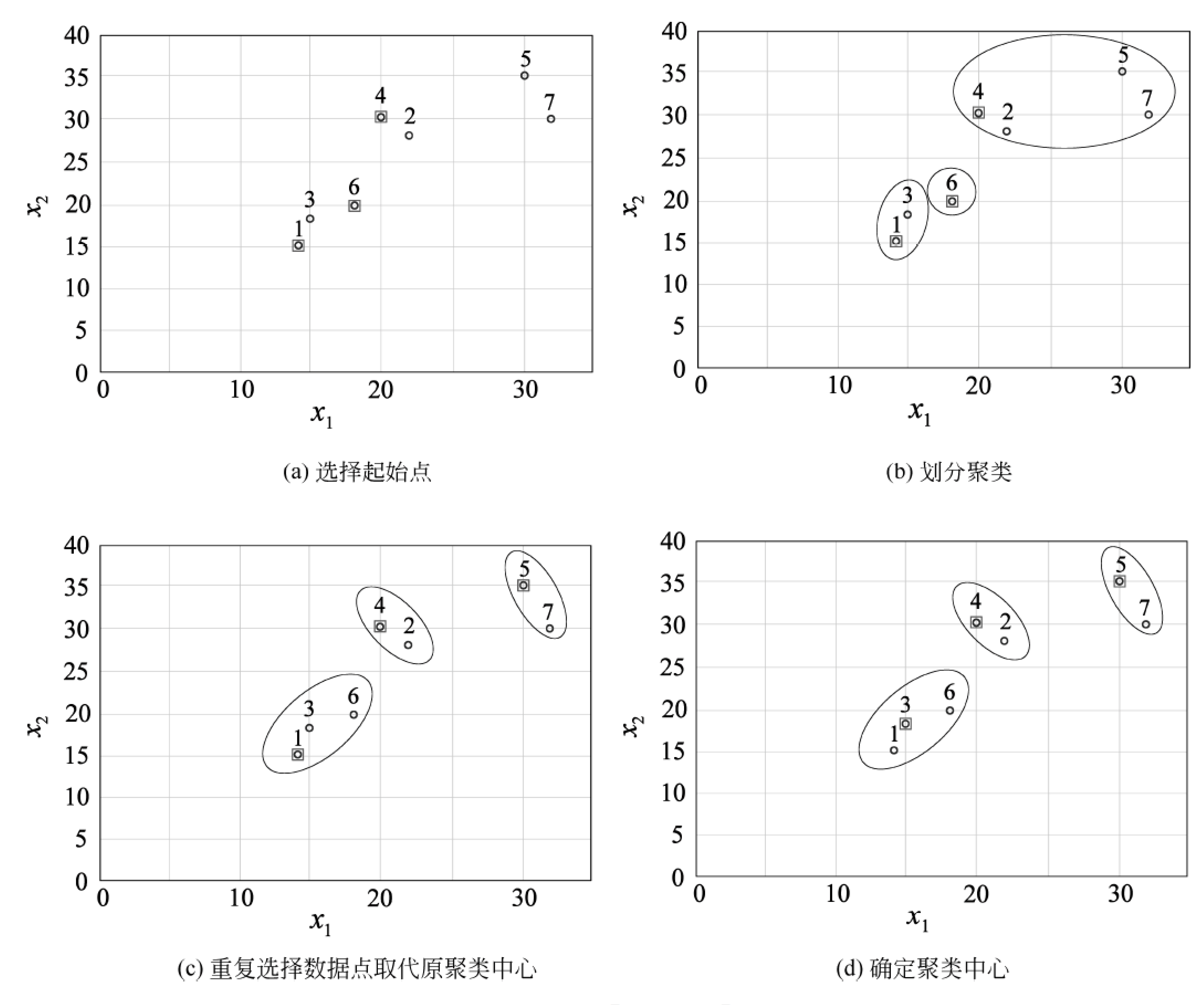
表6.11　K中心点法计算过程2（$K=3$）

序号	与聚类A(1)的相异度	与聚类B(4)的相异度	与聚类C(5)的相异度	最小相异度	分配的聚类
1	0	261	656	0	A
2	233	8	113	8	B
3	10	169	514	10	A
4	261	0	125	0	B
5	656	125	0	0	C
6	41	104	369	41	A
7	549	144	29	29	C

表 6.12　K 中心点法计算过程 3(K=3)

序号	与聚类 A(3)的相异度	与聚类 B(4)的相异度	与聚类 C(5)的相异度	最小相异度	分配的聚类
1	10	261	656	10	A
2	149	8	113	8	B
3	0	169	514	0	A
4	169	0	125	0	B
5	514	125	0	0	C
6	13	104	369	13	A
7	433	144	29	29	C

以[范例 6.1]为例,假设 K 中心法的起始聚类数 K 设为 3,以欧式距离平方作为衡量相似度的依据,先随机选取数据 1、4、6 作为聚类中心,如图 6.7(a),根据所计算的相似度,将数据归类到最近的聚类,形成新的聚类,如图 6.7(b);接着再任选一非聚类数据点,假设以数据点 5 取代聚类 C 的中心数据点 6,如图 6.7(c),再计算总距离改变量 $S=88-287=-199<0$,所以将数据点 5 作为聚类 C 的中心,如此不断重复选择数据点取代原有中心,直至假设最后选择了 3、4、5 作为聚类中心,且聚类中心也不再变动为止,如图 6.7(d)。



(a) 选择起始点　(b) 划分聚类　(c) 重复选择数据点取代原聚类中心　(d) 确定聚类中心

图 6.7　K 中心点法对[范例 6.1]的分群过程

当数据存在噪声与异常值时，K 中心点法比 K 平均法有稳定的分群结果，较不易受到异常值的影响而产生偏差。但当数据点与聚类数目增加时，K 中心点法将需要大量的计算成本，因此有许多算法针对 PAM 算法进行修改以适用于大型数据，如 CLARA(clustering large applications)算法(Kaufman & Rousseeuw，1990)。

6.4　以密度为基础的分群算法

层次聚类分析法和划分聚类分析法都是以数据或聚类间的距离作为分群依据，因此当数据的群聚形状非近似球状时，可能会产生分析误差。**基于密度的聚类方法(density-based clustering)**可处理不同大小、形状聚类的方法，如图 6.8，以密度为导向的分群算法是将密度较高的数据分为一群，未被分配至任一聚类的数据，则会被视为噪声。因此，不但可以针对任意形状的数据分布进行聚类划分，也可以用来过滤异常值与噪声数据。

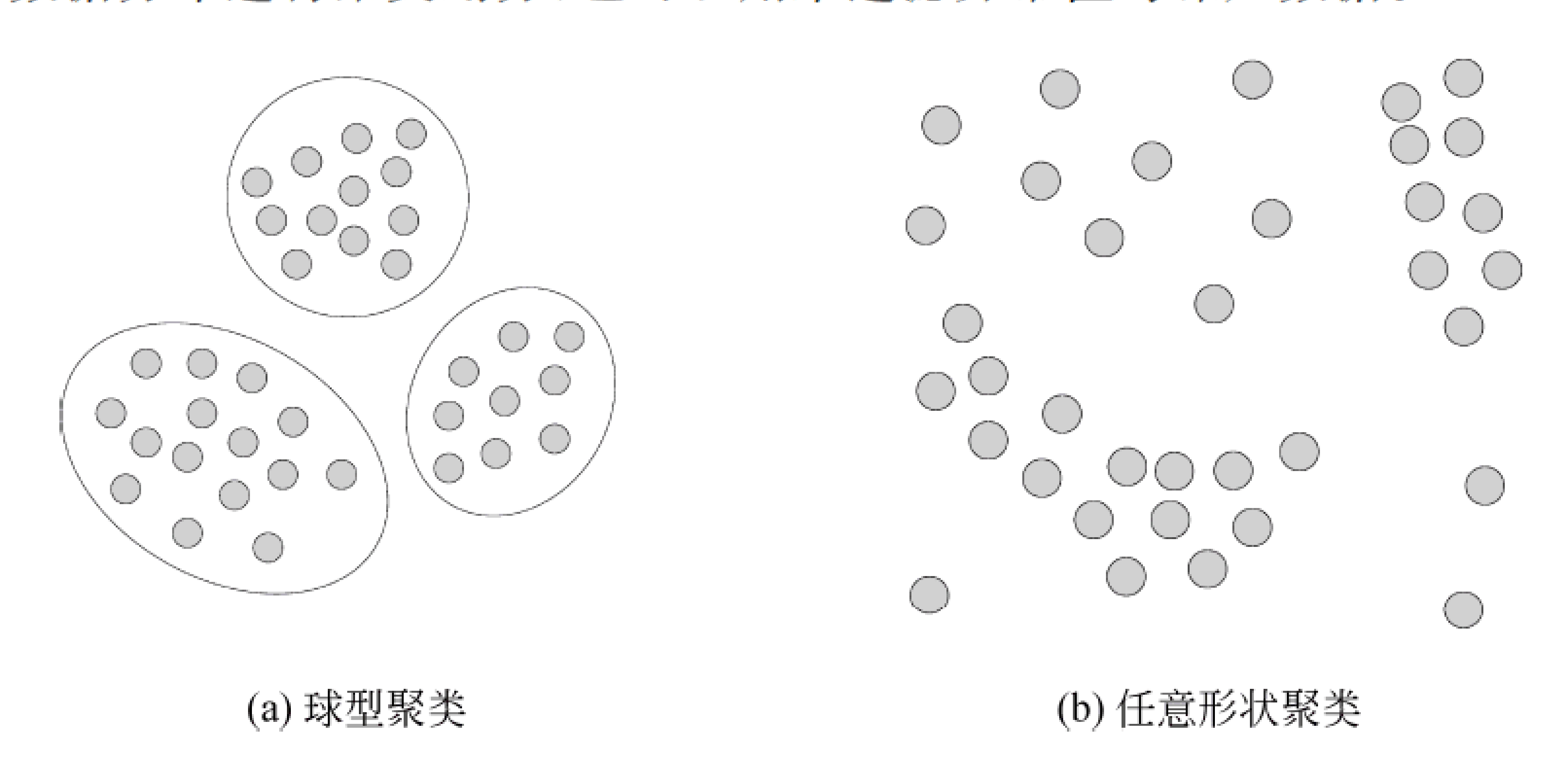


图 6.8　密度概念图

DBSCAN(density-based spatial clustering of applications with noise)是一种基于密度的分群方法(Ester *et al.*，1996)，DBSCAN 主要是判断数据点间的密度是否为密集，高密度的定义为在设定的半径范围参数(ε，Eps)内，所涵盖数据的最小数据数目是否有达到所设定的门槛值(the minimum number of points，MinPts)，若没有达到门槛值，则表示此范围内的数据点不够密集，因此并不需特别划分为一群，反之，则可将数据点聚集成一聚类。不同的聚类间更可利用递移的关系，将较小的聚类聚集成较大的聚类。因此，DBSCAN 可以找到数据点为任意形状分布的聚类。

首先，DBSCAN 可能会出现的数据点种类，可分为三种：

(1) **核心点(core)**：若一个数据点在所定义的半径范围内超过所要求的数据点密度(MinPts)，则此数据点即称为核心点。在图 6.9 中，假设 MinPts 为 4，在设定的半径范围内，数据点 Q、数据点 R 即为核心点。

(2) **境内点(border)**：即落在核心点半径范围内的点。在图 6.9 中，数据点 O 即为境内点。

(3) **噪声点(noise)**：不属于核心点或境内点的数据称作噪声点。在图 6.9 中，数据点 P 即为噪声点。

为了衡量数据点之间的疏密程度，DBSCAN 利用半径范围决定数据点的半径距离，并

利用数据点密度决定聚类内最小数据点数以判断数据点的类型与聚类的结果。DBSCAN 算法的相关定义如下(Ester *et al.*,1996):

- 在设定半径长度内的区域,称为该数据点的 Eps-邻近区域,例如图 6.9 中虚线圆圈的范围。
- 若数据点 S 在核心点 T 的半径范围内,则称数据点 S 从核心点 T 是直接密度可达的(directly density-reachable),当数据点 S 不是核心点 T 时,就无法说数据点 T 从数据点 S 是直接密度可达。如图 6.9 中数据点 Q 与数据点 R 为核心点,则数据点 O 从数据点 Q 为直接密度可达,数据点 Q 从数据点 R 为直接密度可达。
- 若数据点 S 可由点 T_1 直接密度可达,数据点 T_1 可由数据点 T_2 直接密度可达,也就是说当 T_{i-1} 可由数据点 T_i 直接密度可达,则称数据点 S 从数据点 T_i 密度可达。但由于数据点 S 不一定是核心点,所以数据点 T_i 从数据点 S 不一定密度可达。如图 6.9 中数据点 O 从数据点 R 为密度可达,但数据点 R 从数据点 O 不是密度可达,因为数据点 O 不是核心点。
- 若数据点 T_2 与数据点 S 从数据点 T_1 皆是密度可达,则称数据点 T_2 与数据点 S 为密度相连的(density-connected)。如图 6.9 中数据点 U、数据点 W、数据点 V 的关系,数据点 U 与数据点 V 从数据点 W 皆为密度可达,因此数据点 U 与数据点 V 为密度相连。
- 若数据点 T 属于聚类 K,且数据点 S 由数据点 T 密度可达,则数据点 S 也属于聚类 K。另外,在同一聚类内的数据点必为密度相连。如图 6.9 中数据点 U 与数据点 V 为密度相连,因此数据点 U 与数据点 V 属于同一聚类。
- 无法归属到任一聚类的数据点将视为噪声点。

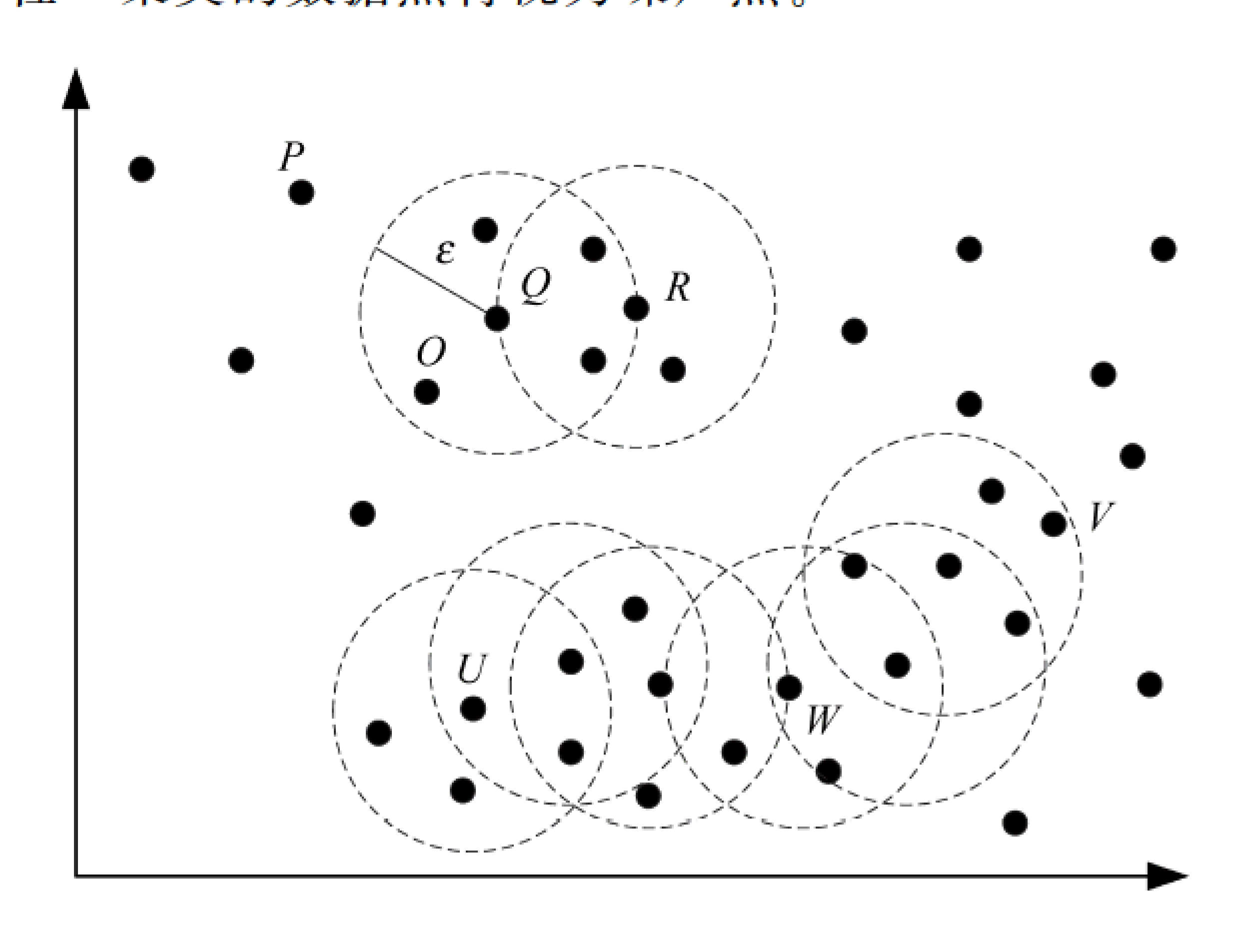


图 6.9 DBSCAN 数据点定义示例

DBSCAN 算法说明如下:首先,判断数据点是否为核心点,接着以核心点为中心,将 Eps-邻近区域的所有境内点合并为一聚类,接下来选择其中一个核心点,并找寻以此核心点密度可达的数据点,若扩张到其他核心点的聚类,则将两聚类合并为一个新的大聚类,若该聚类没有再发现新的核心点,则重新搜索新的核心点,直到所有核心点均被计算过为止。

相对于常见的 K 平均法或层次算法,虽然 DBSCAN 对于有噪声和数据分布为任意形状的数据有较佳的分群结果,但却需要决定适当参数 Eps 与 MinPts;若半径设定得过大,聚

类结果可能会过于粗略，但若半径设得太小，则可能会得到过多的聚类。一般来说，用户可借由重复测试不同的参数组合以找到较为适当的分群结果。然而，当聚类间有不同密度时，由于密度设定的不同会造成聚类的错分，并不建议使用 DBSCAN 算法。

6.5　以模式为基础的分群算法

6.5.1　期望最大化算法

若原始数据是由几个概率分布模型所组成，每个概率模型代表一个群组，在选择 k 个概率密度分布所组成的混合密度模型(mixture density model)下，通过估计这些概率模型，则可计算数据对各概率模型的个体隶属概率(membership probability)，每笔数据根据其最大的隶属概率指派到特定群组则可得到分群结果，由于每笔数据均可能指派到各群组，群体间并没有明确的边界。

假设有 k 个分布模型与 n 个数据点 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，$\Theta$ 为 k 个分布的参数空间 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_k\}$，$\Psi=\{\psi_1,\psi_2,\cdots,\psi_k\}$ 为分布模型占全部数据的比例，$P(x_i|\theta_j)$ 表示第 i 个对象属于第 j 个分布模型的概率，$1\leqslant j\leqslant k$，则观察到数据点 x_i 的概率如式(6.19)：

$$P(x_i \mid \Theta,\Psi) = \sum_{j=1}^{k}\psi_j P(x_i \mid \theta_j) \tag{6.19}$$

假设各数据点为独立，则整个数据集合 $\boldsymbol{X}$ 的概率为个别数据点的概率乘积和：

$$P(\boldsymbol{X} \mid \Theta) = \prod_{i=1}^{n} P(x_i \mid \Theta,\Psi) = \prod_{i=1}^{n}\sum_{j=1}^{k}\psi_j P(x_i \mid \theta_j) \tag{6.20}$$

其中，$P(\boldsymbol{X}|\Theta)$ 代表 n 笔数据的似然函数，若第 j 个分布是高斯分布(Gaussian distribution)，观察值 x_i 来自第 j 个分布的条件概率为 $P(x_i|\theta_j)=\frac{1}{\sqrt{2\pi}\sigma_j}\mathrm{e}^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}}$，$\theta_j=(\mu_j,\sigma_j)$。

期望最大化(expectation-maximization，EM)是通过不断估计模型参数以得到最大似然函数的方法(Dempster *et al.*，1977)。EM 算法先决定 k 组模型的参数，接着计算每个数据点属于每个分布的概率，最后使用这些概率来重新计算模型参数的新估计值，以使得似然值最大化，并且不断迭代改善估计值。EM 算法步骤如下：

(1) 选择分群个数 k，以及 k 组分布模型的参数。

(2) **期望步骤(expectation step)**：已得第 t 次递归的参数估计($\Theta^{(t)}$，$\Psi^{(t)}$)下，计算数据点 x_i 会隶属于聚类 C_j 的概率 $P(C_j|x_i,\Theta^{(t)},\Psi^{(t)})$ 如下：

$$P(C_j \mid x_i,\Theta^{(t)},\Psi^{(t)}) = \frac{\psi_j^{(t)}P(x_i \mid \theta_j^{(t)})}{\sum_{l=1}^{k}\psi_l^{(t)}P(x_i \mid \theta_l^{(t)})} \tag{6.21}$$

数据点隶属于聚类 C_j 概率的计算过程是依据贝叶斯理论(Bayes' theorem)的应用，详细内容请见第 7 章。根据式(6.21)，可得母体参数的期望对数概似函数：

$$\begin{aligned}\boldsymbol{Q}(\Theta \mid X,\Theta^{(t)},\Psi^{(t)}) &= E(\log L(\Theta;X,\Theta^{(t)},\Psi^{(t)}))\\ &= \sum_{i=1}^{n}\sum_{j=1}^{k}P(C_j \mid x_i,\Theta^{(t)},\Psi^{(t)})\psi_j P(x_i \mid \theta_j)\end{aligned} \tag{6.22}$$

(3) **最大化步骤(maximization step)**:使得期望对数似然函数式(6.22)最大的参数估计 $\Theta^{(t+1)}$ 与 $\Psi^{(t+1)}$ 即为所求。

(4) 固定递归次数或估计的模型参数收敛后停止。

EM 与 K 平均法均需事先决定群组数,不同的地方在于 K 平均法递归地估计群中心,将数据点指派至距离最近的聚类;EM 法递归地估计母体参数,将数据点指派至隶属概率最大的聚类。

EM 算法具有容易应用、不受遗漏值的影响(Dempster *et al.*,1977)。此外,由于考虑数据来自不同分布,EM 算法的应用上可以找到具有不同大小与非球状的聚类,例如以高斯分布为基础下,可找到椭圆形的聚类。

6.5.2 自组织映射图网络

自组织映射图网络是聚类分析与数据可视化常用的方法之一。SOM 的好处在于能够处理大量高维度的多变量数据,且同时保留数据所含信息。借由向量量化与向量投影,将多维度的数据映像到二维的拓扑坐标上,并以可视化的方式呈现,辅助对聚类结果的解释。详细内容请参见第 5 章人工神经网络。

6.6 R 语言与聚类分析

本节应用 1973 年美国 50 个州的犯罪率调查统计数据(McNeil,1977),说明如何利用层次聚类法与 K 平均法进行分群。本数据已内建在 R 语言的基础函数库中,共包含 4 个属性与 50 笔观察值,如表 6.13 所示。

表 6.13 聚类分析范例数据

编号	属性名称	属 性 说 明	数据尺度	属性值
1	Murder	每 10 万人中因谋杀被捕的人数	连续	[0.8,17.4]
2	Assault	每 10 万人中因暴力袭击被捕的人数	连续	[45,337]
3	UrbanPop	都市人口比例	连续	[32,91]
4	Rape	每 10 万人中因抢劫被捕的人数	连续	[7.3,46]

层次聚类法与 K 平均法均包含在 R 语言内建的函数库,可通过 **dist** 函数将原始数据转换为以距离为基础的相似度矩阵,并可选择距离函数为欧氏距离、曼哈顿距离、闵氏距离等,进一步通过 **hclust** 函数进行层次集群。同样,用户可选择单一连结法、完全连结法、平均连结法、中心点连结法、沃德法等不同聚集方法。下列程序以欧式距离、通过单一连结法建立阶层聚类为例。

```
data(USArrests)
distance<-dist(USArrests,method="euclidean")
#method 可指定"euclidean","manhattan","minkowski"
hc<-hclust(distance,method="single")
#method 可指定"single","complete","average","centroid","ward"
```

```
plot(hc,hang=-1)
```

图 6.10 与图 6.11 分别为以单一连结法与完全连结法所画出来的阶层聚类图。以单一连结法进行阶层分群容易造成群间的邻近程度较接近，不容易看出观测值间的分群结果；反观以完全连结法进行阶层分群会使得群间的邻近程度拉开，对于聚类个数的决定较容易由图形上检查，以图 6.11 为例，可将数据分为 3 群。

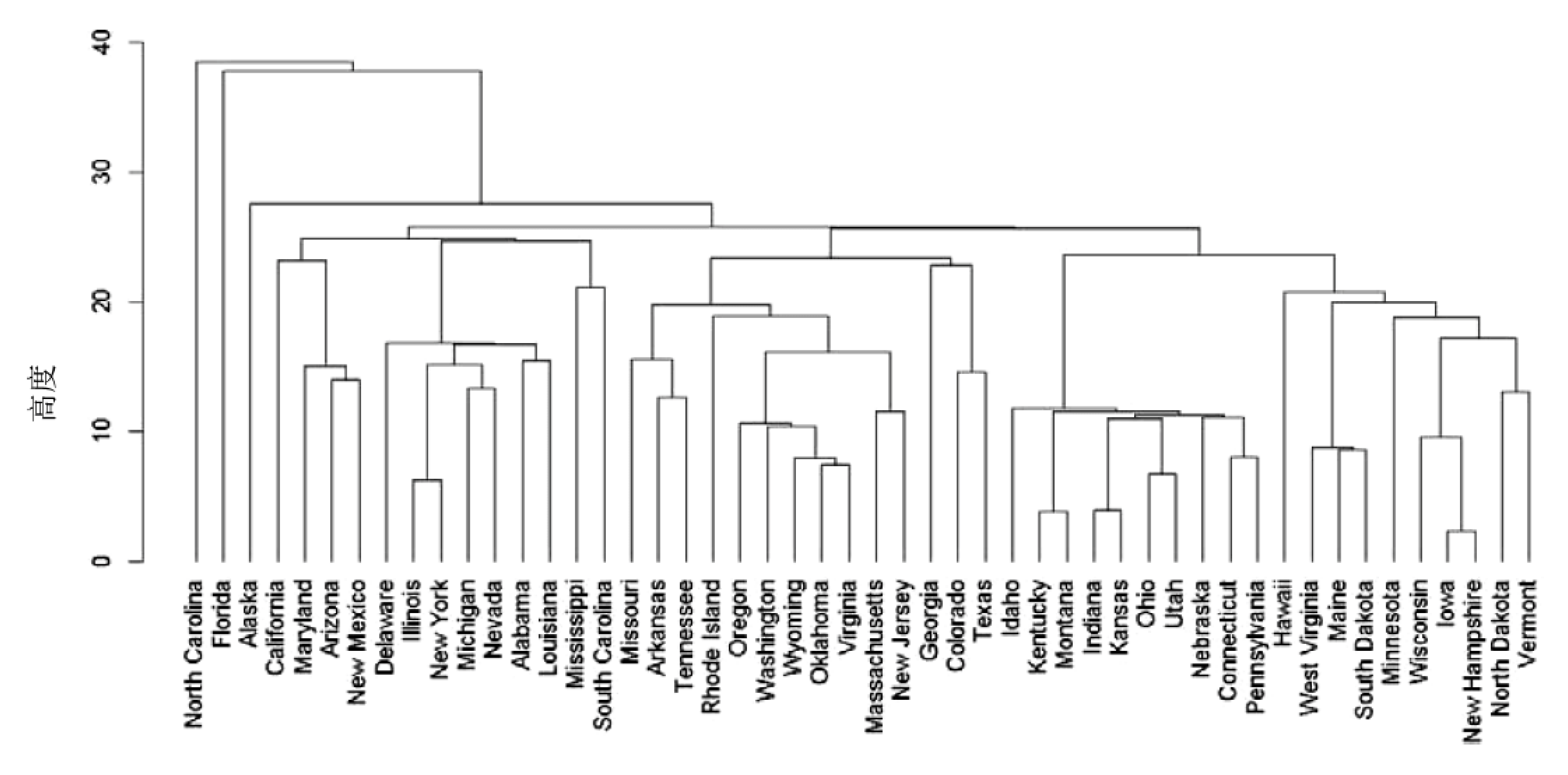


图 6.10　单一连结法层次聚类图

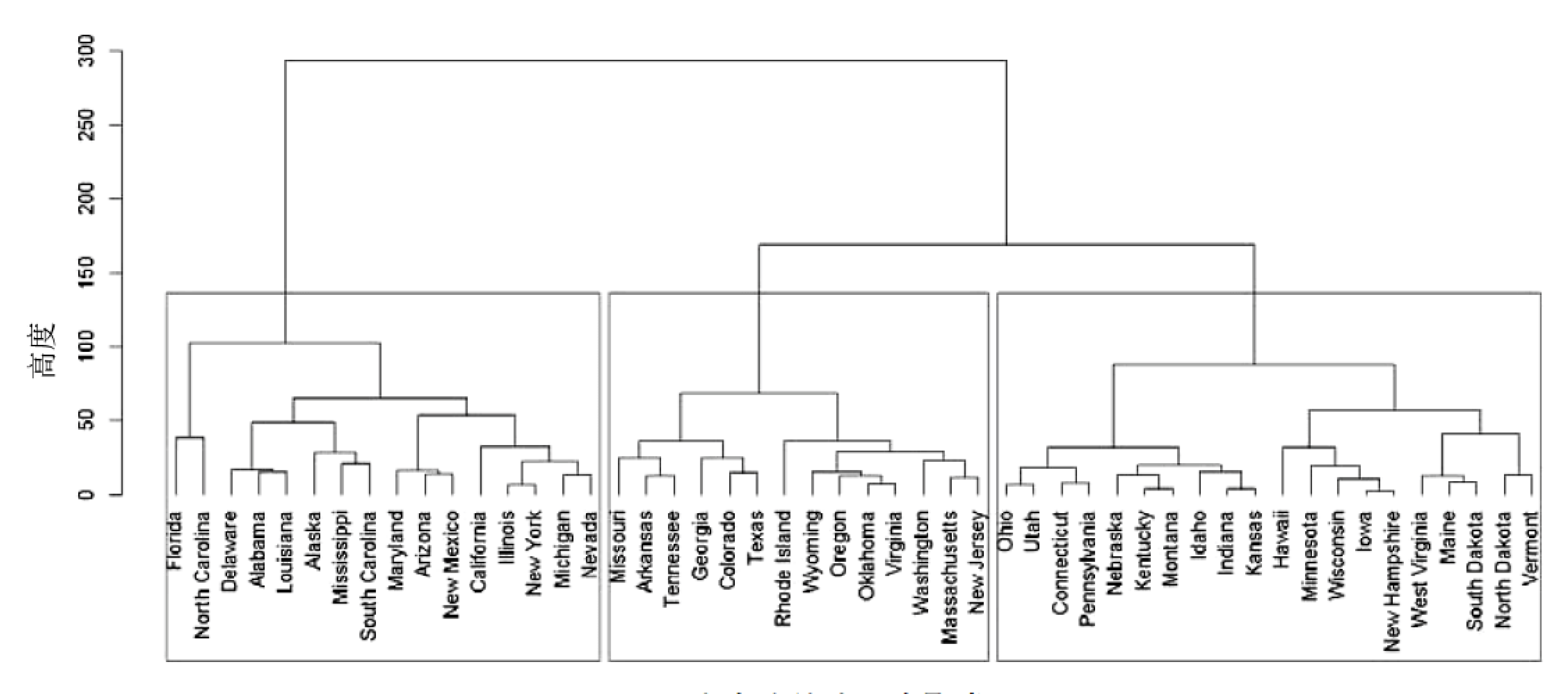


图 6.11　完全连结法层次聚类图

kmeans 函数可用于进行 K 平均法聚类。由于 K 平均法需先给定分群数，在此可先利用层次聚类分析得到适当的分群数，若选择聚类个数为 3 群，接着利用以下指令产生分群结果，如表 6.14 所示，此结果与层次聚类分析的结果一致(图 6.11)。

```
kmeans(USArrests,centers=3)$cluster
```

表 6.14 K 平均法聚类分析结果

聚类一		聚类二		聚类三	
Connecticut	Nebraska	Arkansas	Texas	Alabama	Mississippi
Hawaii	New Hampshire	Colorado	Virginia	Alaska	Nevada
Idaho	North Dakota	Georgia	Washington	Arizona	New Mexico
Indiana	Ohio	Massachusetts	Wyoming	California	New York
Iowa	Pennsylvania	Missouri		Delaware	North Carolina
Kansas	South Dakota	New Jersey		Florida	South Carolina
Kentucky	Utah	Oklahoma		Illinois	
Maine	Vermont	Oregon		Louisiana	
Minnesota	West Virginia	Rhode Island		Maryland	
Montana	Wisconsin	Tennessee		Michigan	

6.7 应用实例——黄光机台聚类分析

6.7.1 案例简介

为了提高生产效率与维持良率,半导体厂往往会将同样类型或表现相近的机台划分为同一聚类,作为互相备援的机台。在同样制程中可能会有数台不同特性的机台。以黄光制程为例,不同曝光机台间的覆盖误差(overlay error)特性也不尽相同,覆盖误差即前一层曝光成像图案位置与后一层曝光成像图案位置的位移误差,覆盖误差必须控制在可被容忍的误差范围内,才不会影响良率(Chien *et al*.,2003)。

黄光制程是半导体制程的瓶颈制程之一,瓶颈机台的利用率往往会由于等待时间过长造成产能下降,考虑到产出因素及避免等待时间的浪费,往往无法在同一机台上完成所有的晶圆曝光程序。因此,为了增加曝光机台的产出,避免不同层间严重的对准不良造成良率的损失,必须选择覆盖误差特性近似的机台作为前后层的作业或备用机台,如此一来,当机台必须维护保养时,属于同一聚类的机台即可马上替补,以维持晶圆产出并避免影响到良率。然而,工程师往往凭借对机台的了解与过去经验作为指定配对机台的依据,而缺乏自动化判断分群的机制。曝光机台覆盖误差特征会随着制程、时间而改变或飘移,仅依靠经验法则可能会因没有掌握到目前各曝光机台的状况而造成良率损失。随着曝光机台功能的进步,覆盖误差的容差界限已越来越紧缩,在考虑降低成本与提高产能的要务下,机台聚类以及备用机台的选择已成为极重要的决策问题(Chien & Hsu,2006)。

通过覆盖误差模式可将覆盖误差(d_{x+X},d_{y+Y})解构为系统性与非系统性覆盖误差。考虑实际可被机台补偿的覆盖误差因子,系统性覆盖误差根据产生的来源包括曝光区域内,平移误差(T_{x+X},T_{y+Y})、intrafield 的放大误差(M'_x,M'_y)、intrafield 的旋转误差(R_x,R_y)、interfield 的放大倍率误差(S_X,S_Y)、interfield 的旋转误差(R_X,R_Y)。非系统性覆盖误差指的是无法被机台校正补偿之覆盖误差,如透镜指纹造成的像差,或者是随机误差等无法被机

台补偿校正的误差，在模式中代表的参数为残差。本案例套用实际半导体步进机覆盖误差模式（Chien *et al.*，2003），以最小二乘法估计各覆盖误差因子与残差，以作为后续机台在系统性误差相似度的比较依据。

$$d_{x+X} = T_{x+X} + S_X X - (N+\theta)Y + M'_x x - R_x y + \varepsilon_{x+X} \tag{6.23}$$

$$d_{y+Y} = T_{y+Y} + S_Y Y - (\theta - N)X + M'_y y - R_y x + \varepsilon_{y+Y} \tag{6.24}$$

其中，$N=(R_X-R_Y)/2$，$\theta=(R_X+R_Y)/2$。

本案例采用两阶段分群法比较系统性覆盖误差相似度，第一阶段是使用沃德法与 RMSSTD(root mean square standard deviation)、R-square、SPR (semi-partial R-square)三个指标(Subhash，1996)找出适当分群个数，第二阶段则根据沃德法所得出的群数，再用 K 平均法重新分群，而被归类于同一群的机台，表示该群机台具有相似的系统性误差特性。

6.7.2　验证两阶段分群算法

案例中选择某 DRAM 厂黄光制程中 10 台步进式曝光机的数据进行实证，一片晶圆上量测 5 个曝光区域，一个曝光区域内测量 4 个覆盖误差。利用最小二乘法估计各覆盖误差因子的参数值，各机台估计的覆盖误差因子如表 6.15 所列，在衡量两机台间在系统性覆盖误差的相似度采用欧氏距离平方。

接着以 10 台曝光机的系统性覆盖误差因子为分群特征变量进行两阶段聚类分析，首先根据沃德法决定聚类个数，第一阶段沃德法分群结果如图 6.12 所示，各分群评估指标由图 6.13，可发现分群个数由 4 群缩减为 3 群时，RMSSTD 与 R-square 增减的幅度较大，表示由 4 个群体合并为 3 个群体时，群内机台的相似性显著降低，群间的相异性也随之降低。且 SPR 增大，表示结合成 3 个群体时，群内机台相似性损失的比例增大。因此，综合考虑以上 4 个分群指标，重新规定聚类个数 $K=4$。再用 K 平均法重新对 10 台曝光机分群，其分群结果如表 6.16 所列。

表 6.15　不同机台的覆盖误差因子

序号	T_{x+X}	T_{y+Y}	S_X	S_Y	R_X	R_Y	M'_x	M'_y	R_x	R_y
1	−0.010	−0.047	−0.115	0.024	−0.078	−0.091	2.304	6.416	−3.422	3.281
2	−0.018	−0.032	0.091	0.392	−0.143	0.043	1.828	−0.455	−2.209	3.441
3	0.033	0.003	0.126	0.856	−0.300	−0.120	−2.146	0.935	−1.605	1.501
4	0.019	−0.007	1.142	1.260	−0.765	0.236	−0.256	−0.331	−3.633	1.369
5	−0.056	−0.056	0.364	0.613	−0.104	−0.075	5.527	4.625	−3.662	4.091
6	0.032	0.017	0.284	0.688	−0.314	0.028	1.027	2.387	−3.435	1.719
7	0.055	0.016	−0.272	0.015	−0.250	−0.012	−1.713	0.285	−4.429	−0.564
8	−0.012	0.113	0.339	0.635	−0.472	0.168	1.731	2.595	−3.402	−4.146
9	−0.021	−0.032	−0.203	0.124	−0.059	−0.060	2.879	3.096	−5.775	4.348
10	0.075	0.014	−0.128	0.045	−0.109	−0.142	−0.856	−2.112	−3.352	−1.305

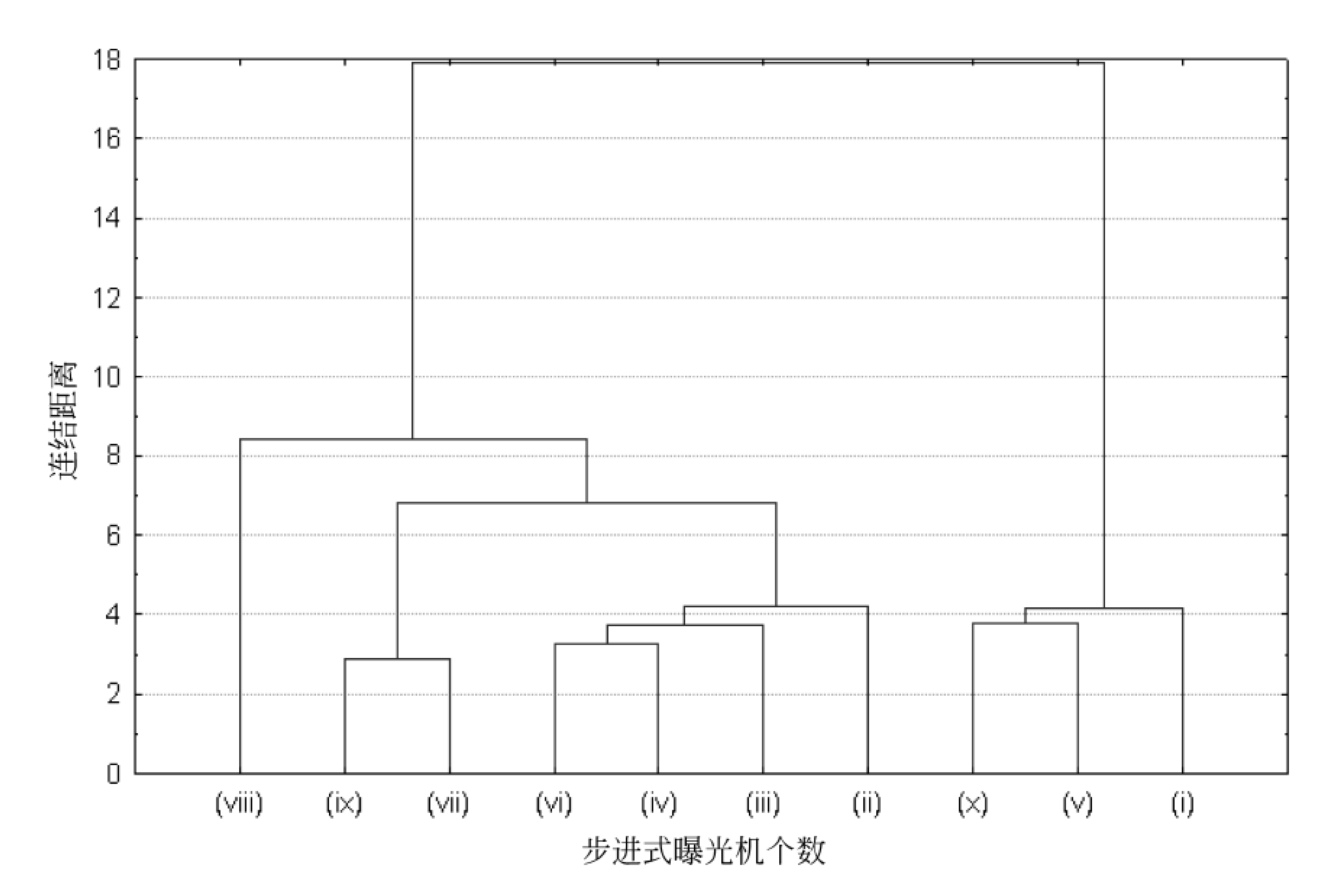


图 6.12 沃德法分群树形图

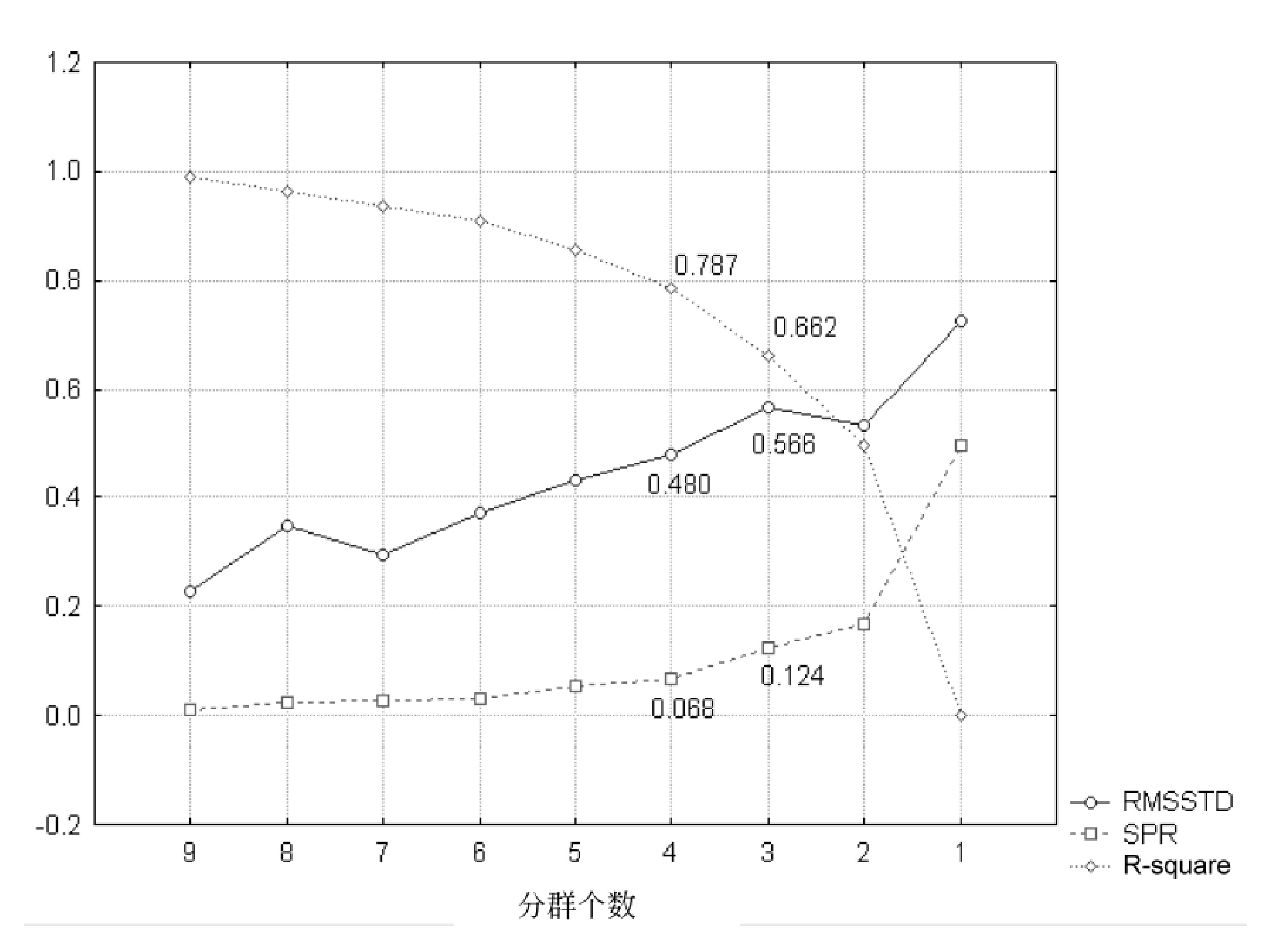


图 6.13 衡量分群个数的指标

表 6.16 *K* 平均法分群结果(*K*=4)

聚类	步进机			
1	(viii)			
2	(ii)	(iii)	(iv)	(vi)
3	(i)	(v)	(ix)	
4	(vii)	(x)		

6.7.3 案例小结

本研究针对半导体黄光制程曝光机台的聚类问题，提出机台分群算法，可协助工程师指派覆盖误差相似的曝光机台，及作为寻找更佳替代机台的决策依据，减少因备用机台安排错误所造成的良率损失。借由实际机台所使用的覆盖误差模式，估计各机台在晶圆上造成的覆盖误差，并运用回归分析中的最小二乘法，计算各误差因子的补偿参数；经由模式依覆盖误差结构分为系统性与非系统性误差两部分；针对系统性误差，比较各曝光机台的覆盖误差因子间的相似程度，找出覆盖误差因子相似的机台聚类。针对非系统性误差，也可利用各曝光机台 X 与 Y 方向残差的相关系数以衡量其相似度。在工程师搜集实际生产所量得的数据后，据此验证所提出的机台分群算法并与工程师讨论，发现结果与工程师指派备用机台的专业经验互相符合。此研究成果除了帮助工程师指派备用机台外，并可将各曝光机台间的相似性结果列入黄光曝光机台日程安排的考虑，作为制造执行系统在生产排程上依据，以兼顾生产产出率与良率。

6.8 结论

本章介绍的聚类分析主要针对相似度的计算、分群算法进行说明，不同聚类分析算法均有适合应用的数据与问题。层次聚类分析利用凝聚或分裂的过程，将相似度较高的个体或聚类合并为同一聚类，对于处理聚类大小差异大、存在异常值的数据结果均较划分聚类分析好，然而缺点是当数据笔数较多、变数维度过高时，需耗费较多的计算时间。划分聚类分析则是先决定聚类个数，根据定义的聚类质量，如聚类内数据变异最小，直接将数据划分至数个没有交集的聚类，并通过反复比较与重新归属以提升所定义的聚类质量。对于找出近似圆形的聚类或希望聚类内数据个数差异不大的情况下会有较佳的结果。以密度为基础的分群算法则可针对任意形状聚类进行分析，且不容易受到噪声数据的影响，缺点是最佳设定参数往往难以设定。当已知聚类的数据分布，则可利用以模式为基础的聚类分析。

分群结果会受所选择的量测尺度和衡量相似度的标准所影响，因此必须格外小心。不管是哪一种衡量方法，都必须配合数据类型与聚类分析算法。例如，单一连结法与完全连结法是使用最大或最小距离来衡量，但这样的方法对于噪声数据或异常值往往会过于敏感，所以可改用平均连结法或中心点连结法来解决。沃德法倾向将聚类切分为几个小群，优点是可自行判断分群个数。层次聚类分析法虽然不需要事先决定群数，但是一旦被分到同一群就无法再分开，划分聚类分析法则是每次均重新计算各群体到中心的距离，所以可弥补层次聚类分析法的缺点。

聚类分析可将数据分成数个子聚类，使得各聚类内个体相似度高，聚类间相似度低。除此之外，更重要的是找出有意义的聚类，也就是说聚类分析结果的好坏必须回到数据的本质上是否可解释。再者，不同聚类所使用的不同目标也会导致不同的结果，举例来说，均方误差和对于衡量非球状的聚类可能是无意义的，在实务应用上，考虑不同数据形态，仍须依赖分群目的而选择对应的聚类分析算法。

聚类分析不仅可独自从一堆数据中探索，找出数据子聚类间的特征，在许多实务问题上也可根据不同分析目的，将聚类分析的结果作为后续分析模式的输入数据。例如当数据维

度众多，可先利用聚类分析找出数据间的特征，再将其提取出有意义的特征与分类算法整合，从中找出影响不同聚类间重要的关联变量。

问题与讨论

1. 以[范例 6.1]为数据，用曼哈顿距离和完全连结法来建立层次聚类分析的树形图。

2. 不同的输入数据是否会对聚类分析的结果造成影响？请举例说明。

3. 试比较层次聚类分析、划分聚类分析、以密度为基础的分群算法、以模式为基础的分群算法间的优缺点。

4. 假设有 3 笔观察值：

A：(10,5,100,23)

B：(12,10,50,40)

C：(8,15,10,20)

(1) 试利用本书中所提及的 3 种距离作为相似度衡量的依据，计算以下数据的相似度。

(2) 该组数据是否需要归一化？

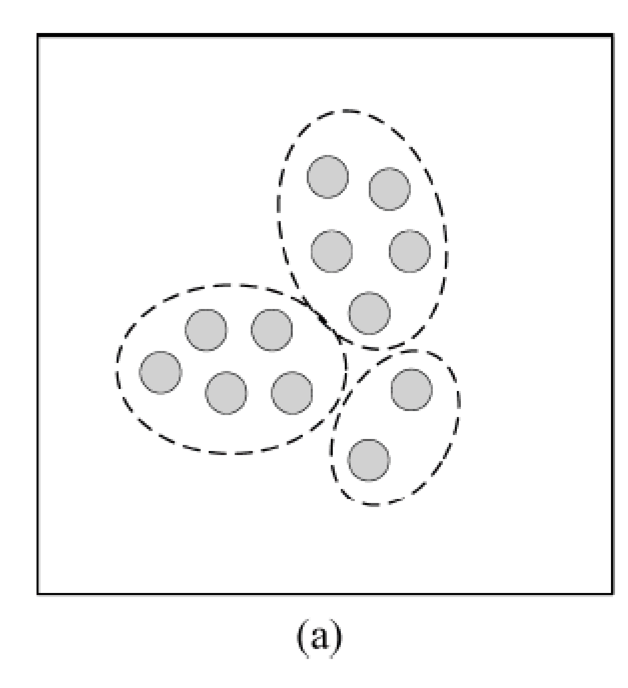
(3) 试利用单一连结法、完全连结法、平均连结法、中心点连结法、沃德法对该组数据进行分群。

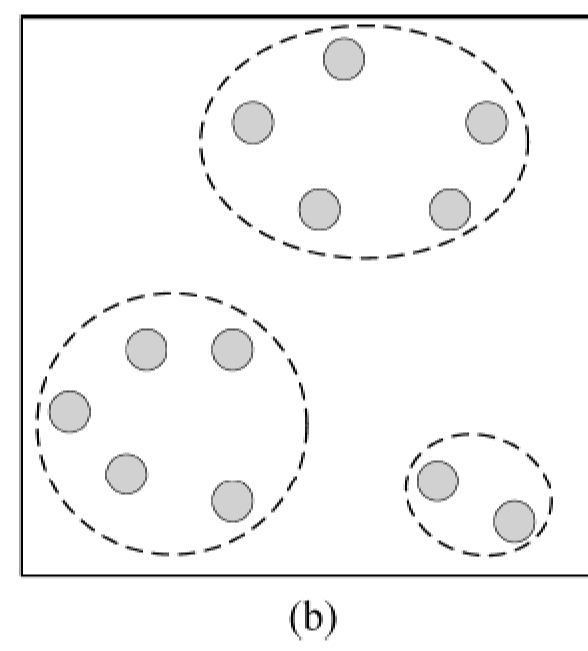
5. 聚类分析是数据挖掘重要的方法之一，请分别举出仅应用聚类分析的案例，以及以聚类分析作为前处理的案例。

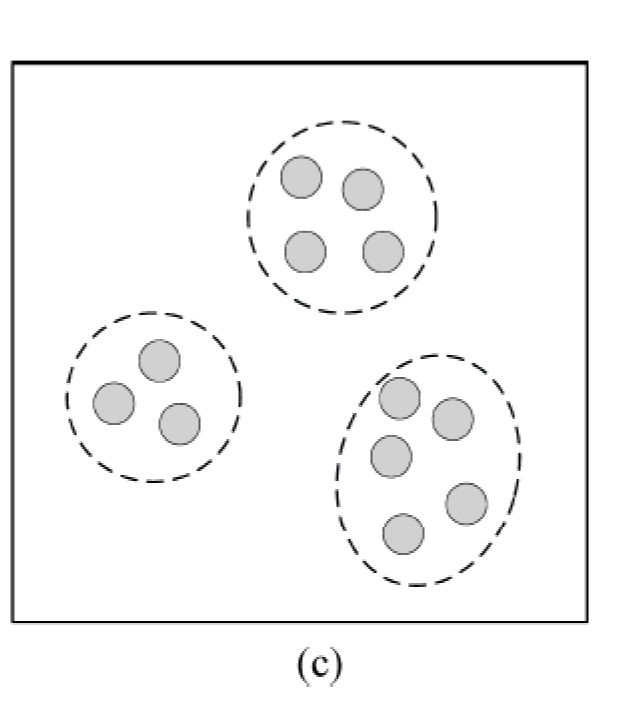
6. 聚类分析结果的好坏往往会利用二维数据来呈现，针对更高维度的数据该如何检查分群结果的好坏，试以三维数据为例，说明你的想法。

7. 该如何说明一个聚类分析的结果是有意义的？

8. 下图为同一笔数据经由三种聚类分析算法得到的结果，请问哪一个聚类分析的结果较佳？试说明你的想法。


(a)


(b)


(c)

9. 给定以下 5 笔数据的相似度矩阵：

	A	*B*	*C*	*D*	*E*
A	1.00	0.20	0.31	0.55	0.80
B	0.20	1.00	0.66	0.35	0.98
C	0.31	0.66	1.00	0.44	0.85
D	0.55	0.35	0.44	1.00	0.70
E	0.80	0.98	0.85	0.70	1.00

（1）利用完全连结法对这5笔数据进行层次聚类分析，并画出对应的树形图。

（2）试决定会有多少聚类个数，决定的相似度门槛是多少？

（3）假设相似度门槛值是0.4，设定MinPts至少大于2，请找出在数据表中的核心点、境内点、噪声点。

10. 试说明以K平均法及K中心点法将表6.2的数据分为三群时，起始聚类中心所造成的影响。

11. 若表6.2的第一笔观察值误植为(1400,15)，试以单一连结法及中心点连结法得出层次聚类。

第 7 章

朴素贝叶斯分类法与贝叶斯网络

贝叶斯分类(Bayesian classifier)借由数据中分析属性与反应变量之间的概率模型，根据贝叶斯定理(Bayes' theorem)来更新信息以推理判断样本数据归属的类别，作为分类和推论的依据，常用的方法有朴素贝叶斯分类法(naive Bayesian classifier)及贝叶斯网络分类法(Bayesian network classifier，简称贝叶斯网络)。由于并非所有的事件都有大量的历史数据或可以重复实验，因此面对没有经验、可参考的信息过少或者没有频率概率存在的情况，贝叶斯网络亦可采用主观概率(subjective probability)，亦即将认为该事件是否会发生的置信度(degree of belief)的主观判断转为主观概率。以下先介绍贝叶斯定理，再依序介绍朴素贝叶斯分类法、贝叶斯网络以及案例分析。

7.1 贝叶斯定理

贝叶斯定理是根据新的信息将先验概率(prior probability)修正为验后概率(posterior probability)的过程。**条件概率(conditional probability)**是根据某一事件发生的情况下，估计另一事件发生的概率，所以验后概率是给定新的信息或证据下的条件概率。贝叶斯理论的主要概念为，一开始不知道目标事件 $\tilde{\theta}$ 的真实状态，但知道 $\tilde{\theta}$ 服从一个概率分布 $P(\tilde{\theta})$，称为先验概率。当得到新的样本信息或证据 E 后，可以根据式(7.1)贝叶斯定理，更新验后概率 $P(\tilde{\theta}|E)$。

$$P(\tilde{\theta}=\theta_j \mid E)=P(\theta_j \mid E)=\frac{P(\theta_j \cap E)}{P(E)}=\frac{P(E \mid \theta_j)\cdot P(\theta_j)}{\sum_{j=1}^{m} P(E \mid \theta_j)\cdot P(\theta_j)} \tag{7.1}$$

如果 E 为特定事件或证据(evidence)，$\tilde{\theta}=\theta_j$ 为某假设(hypothesis)，则在事件 E 发生的情况下，θ_j 发生的条件概率 $P(\theta_j|E)$可表示如式(7.2)：

$$P(\theta_j \mid E)=\frac{P(\theta_j \cap E)}{P(E)} \tag{7.2}$$

其中，$P(E)$为事件 E 发生的概率，$P(\theta_j \cap E)$代表假设 θ_j 与事件 E 同时发生的概率，其概率又可表示如式(7.3)：

$$P(\theta_j \cap E)=P(\theta_j \mid E)\cdot P(E)=P(E \mid \theta_j)\cdot P(\theta_j) \tag{7.3}$$

若 $\theta_1,\theta_2,\cdots,\theta_m$ 为假设 $\tilde{\theta}$ 在样本空间 S 中的一个分割，且事件 $E \subset S$，$P(\theta_j)\neq 0$，$j=1,2,\cdots,m$，则根据全概率定理，$P(E)$可定义如式(7.4)：

$$P(E)=P(E\mid\theta_1)\cdot P(\theta_1)+P(E\mid\theta_2)\cdot P(\theta_2)+\cdots+P(E\mid\theta_m)\cdot P(\theta_m)$$
$$=\sum_{j=1}^{m}P(E\mid\theta_j)\cdot P(\theta_j) \tag{7.4}$$

在取得的新信息事件 E 下，贝叶斯定理可修正假设 $\tilde{\theta}=\theta_j$ 的先验概率 $P(\theta_j)$ 为验后概率 $P(\theta_j|E)$，如式(7.5)所示：

$$P(\theta_j\mid E)=\frac{P(E\mid\theta_j)\cdot P(\theta_j)}{\sum_{j=1}^{m}P(E\mid\theta_j)\cdot P(\theta_j)} \tag{7.5}$$

似然函数 $P_{\tilde{\theta}}(x)$ 也可以表示为 $P(x|\tilde{\theta})$，其中，$\tilde{\theta}$ 代表一随机变量，虽然形式和条件概率雷同，但含义并不相同。条件概率 $P(\tilde{x}=x_i|\theta_j)$ 是代表在给定 $\theta=\theta_j$ 的条件下，随机变量 $\tilde{x}=x_i$ 的概率有多高，此时 θ_j 并非随机变数而是母体参数，如图 7.1 所示；而似然函数 $P(x_i|\tilde{\theta}=\theta_j)$ 是观察到 x_i 时有多少可能性是来自于随机变量 $\tilde{\theta}=\theta_j$ 的情况，如图 7.2。若一个样本 x_i 在真实状态为 $\tilde{\theta}=\theta_j$ 时被观察到的可能性很高，则样本 x_i 对决策者判断真实状态是否为 θ_j 有很高的信息价值。换言之，$P_{\theta_j}(x_i)$ 越高则决策者观察到样本 x_i 后，对 $\tilde{\theta}=\theta_j$ 的信心(belief)越高。

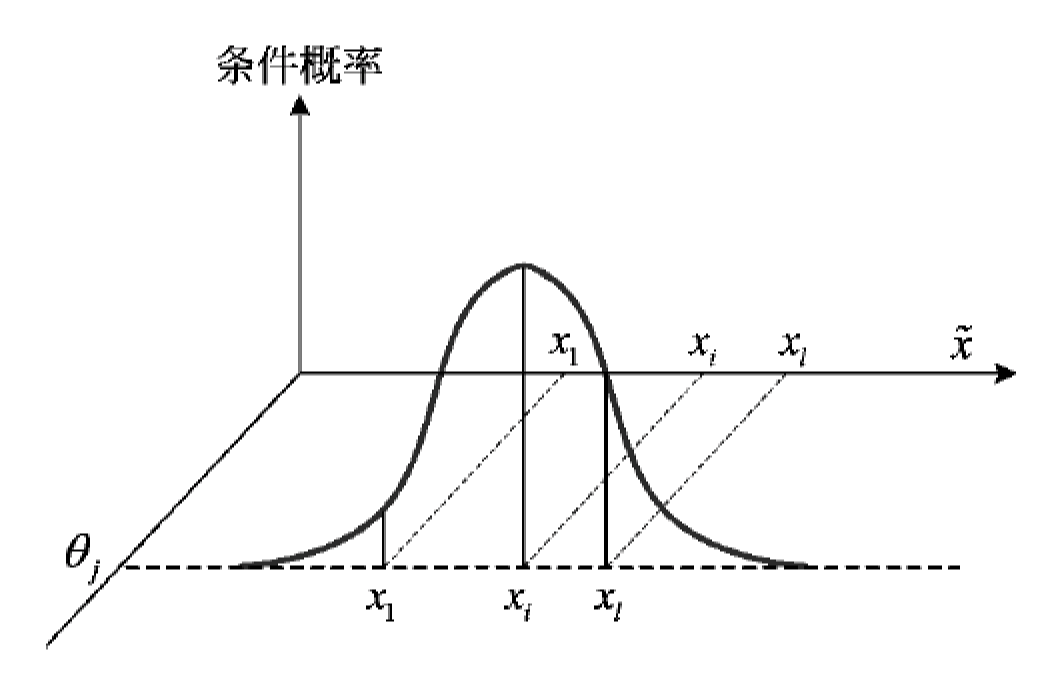


图 7.1　条件概率 $P(\tilde{x}=x_i|\theta_j)$ 示意图

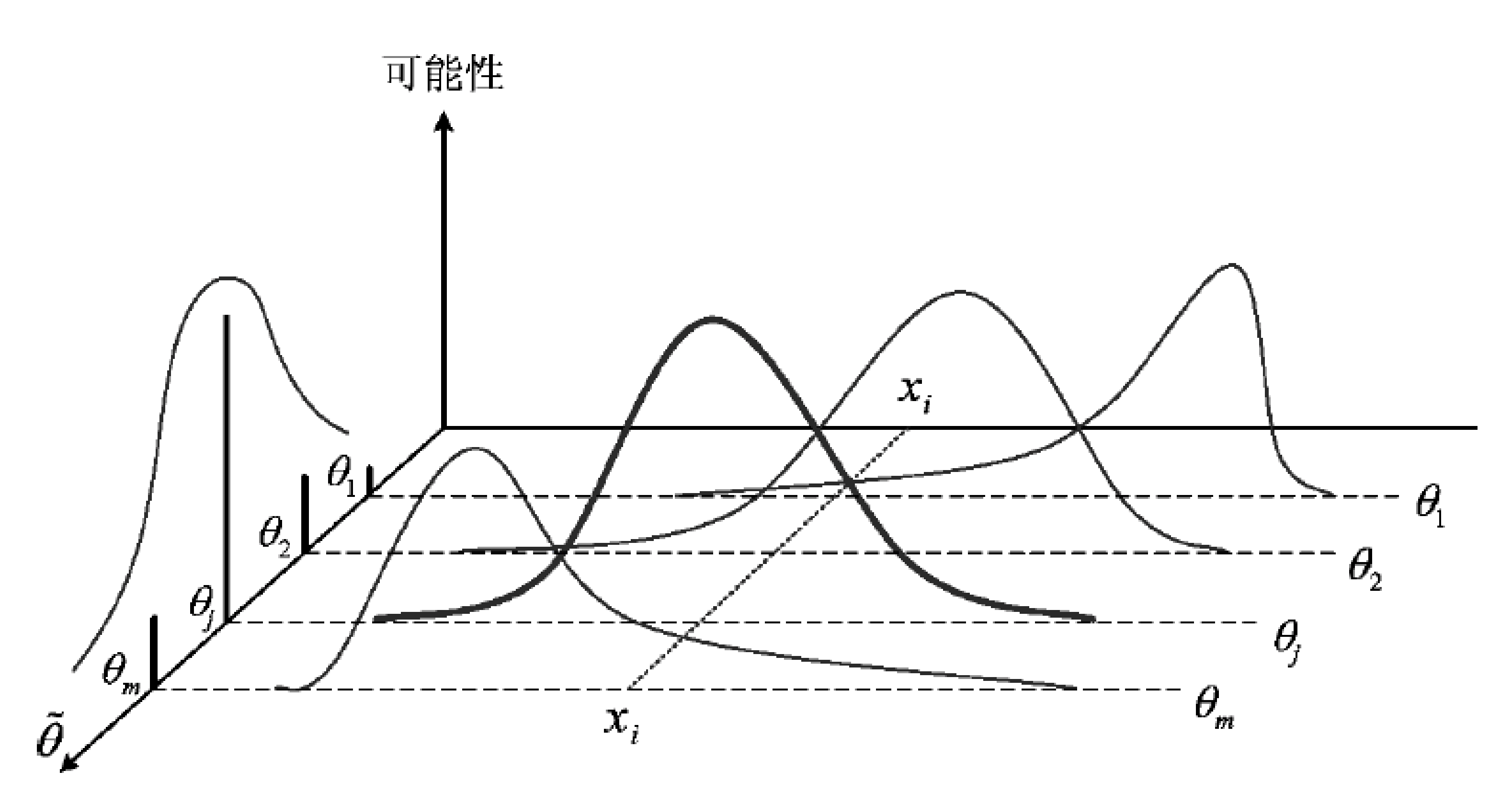


图 7.2　似然函数 $P(x_i|\tilde{\theta}=\theta_j)$ 示意图

简而言之，当参数 $\tilde{\theta}=\theta_j$ 为已知条件时，则 $P(\tilde{x}|\tilde{\theta}=\theta_j)$ 可视为条件概率(其中，$\tilde{x}$ 为随

机变数)；反之，若参数 $\tilde{\theta}$ 为未知时，但已观测变量 $\tilde{x}=x_i$ 的结果，则 $P(x_i|\tilde{\theta}=\theta_j)$ 可视为各可能参数 θ_j 发生的可能性(此时，$\tilde{\theta}$ 为随机变数)。

若 $\Omega=\{\theta_j|j=1,2,\cdots,m\}$，且 $\sum_{j=1}^{m}P(\theta_j)=1$，而 $X=\{x_i|i=1,2,\cdots,l\}$，且 $\sum_{i=1}^{l}P(x_i)=1$。条件概率 $P(\tilde{x}=x_1|\theta_j)$ 表示在 θ_j 的条件下，$\tilde{x}=x_1$ 的概率，因此：

$$P(\tilde{x}=x_1 \mid \theta_j)=\frac{P(x_1 \cap \theta_j)}{\sum_{i=1}^{l}P(x_i \cap \theta_j)} \tag{7.6}$$

[**范例 7.1**] 若某品牌手机主要由 A、B 两家工厂生产，而工厂 A 的生产量为工厂 B 的 4 倍，且已知工厂 A 的良率为 15/16，工厂 B 的良率为 3/4。

$$P(\text{良品} \mid \text{工厂 A 所生产})=\frac{750}{800}$$

$$P(\text{良品} \mid \text{工厂 B 所生产})=\frac{150}{200}$$

由此可求得当检验结果为不良品时，该不良品来自于工厂 A 的可能性有 0.5，来自工厂 B 的可能性有 0.5，计算如下：

$$P(E=\text{工厂 A 所生产} \mid \tilde{\theta}=\text{不良品})=\frac{P(\text{不良品且工厂 A})}{P(\text{不良品且工厂 A})+P(\text{不良品且工厂 B})}$$

$$=\frac{P(\text{不良品} \mid \text{工厂 A})P(\text{工厂 A})}{P(\text{不良品} \mid \text{工厂 A})P(\text{工厂 A})+P(\text{不良品} \mid \text{工厂 B})P(\text{工厂 B})}$$

$$=\frac{\frac{50}{800}\times\frac{800}{1000}}{\frac{50}{800}\times\frac{800}{1000}+\frac{50}{200}\times\frac{200}{1000}}=0.5$$

$$P(E=\text{工厂 B 所生产} \mid \tilde{\theta}=\text{不良品})=\frac{P(\text{不良品且工厂 B})}{P(\text{不良品且工厂 B})+P(\text{不良品且工厂 A})}$$

$$=\frac{P(\text{不良品} \mid \text{工厂 B})P(\text{工厂 B})}{P(\text{不良品} \mid \text{工厂 B})P(\text{工厂 B})+P(\text{不良品} \mid \text{工厂 A})P(\text{工厂 A})}$$

$$=\frac{\frac{50}{200}\times\frac{200}{1000}}{\frac{50}{200}\times\frac{200}{1000}+\frac{50}{800}\times\frac{800}{1000}}=0.5$$

7.2 朴素贝叶斯分类法

朴素贝叶斯分类法又称为单纯贝叶斯分类法，有两项基本假设：①已知各类别的先验概率，常依据专家意见、历史数据或训练数据设定；②给定任一类别下，属性数据相互独立，即属性数据条件独立(conditional independence)。

当预测数据集不包含属性数据时，只能依据先验概率预测观察值属于何种类别。但当预测数据集包含属性数据时，则可依据属性数据建立各分类的条件概率模型，再利用预测数据集的属性数据与贝叶斯定理，算出每笔属性数据属于各分类的验后概率，将属性数据归类

于验后概率最大的类别。朴素贝叶斯分类法亦能进行高维度数据分类，并快速构建可用于分类和预测的数据挖掘模型。

假设一训练数据集包含 n 笔数据，$i=1,2,\cdots,n$，其中，有 m 个类别 $\tilde{\theta}=\{\theta_1,\theta_2,\cdots,\theta_m\}$，其对应先验概率为 $P(\tilde{\theta}=\theta_j)$，$j=1,2,\cdots,m$，定义第 i 笔数据中 k 个属性的观察值为 $\boldsymbol{E}_i=(E_{i1},E_{i2},\cdots,E_{ik})^{\mathrm{T}}$，并令 $\boldsymbol{E}=(\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_n)^{\mathrm{T}}$ 代表训练数据集的所有属性数据。朴素贝叶斯分类法利用最大化各类别的条件概率分布 $P(\boldsymbol{E}|\tilde{\theta}=\theta_j)$ 得到各类别的条件概率模型，再利用数据集的属性数据 $\boldsymbol{E}^*=(E_1^*,E_2^*,\cdots,E_k^*)$ 与贝叶斯定理算出各分类的验后概率：

$$P(\theta_j \mid \boldsymbol{E}^*)=\frac{P(\boldsymbol{E}^* \mid \theta_j)\cdot P(\theta_j)}{P(\boldsymbol{E}^*)},\quad j=1,2,\cdots,m \tag{7.7}$$

其中，$P(\tilde{\theta}=\theta_j)$ 为已知各类别的概率，可由假设①得到，$P(\boldsymbol{E}^*)=\sum_{j=1}^{m}P(\boldsymbol{E}^* \mid \tilde{\theta}=\theta_j)P(\tilde{\theta}=\theta_j)$，表示观察到 $\boldsymbol{E}^*$ 属性的概率。当

$$P(\theta_j \mid \boldsymbol{E}^*)>P(\theta_s \mid \boldsymbol{E}^*),\quad j=1,2,\cdots,m,\ j\neq s \tag{7.8}$$

则推测属性 $\boldsymbol{E}^*$ 应该来自于类别 θ_j。

然而，属性数据常包括不只一个变量 E_l^*，即 $k>1$，$l=1,2,\cdots,k$，利用假设②的条件独立可得

$$\begin{aligned}P(\boldsymbol{E}^* \mid \theta_j)&=P(E_1^*,E_2^*,\cdots,E_k^* \mid \theta_j)\\&=P(E_1^* \mid \theta_j)\cdot P(E_2^* \mid \theta_j)\cdot\cdots\cdot P(E_k^* \mid \theta_j)\\&=\prod_{l=1}^{k}P(E_l^* \mid \theta_j)\end{aligned} \tag{7.9}$$

由式(7.9)可得验后概率如式(7.10)：

$$P(\theta_j \mid \boldsymbol{E}^*)=\frac{\prod_{l=1}^{k}P(E_l^* \mid \theta_j)\cdot P(\theta_j)}{\sum_{j=1}^{m}\prod_{l=1}^{k}P(E_l^* \mid \theta_j)\cdot P(\theta_j)} \tag{7.10}$$

在最大化各类别的条件概率分布 $P(\boldsymbol{E}|\theta_j)$ 时，往往会附加额外的假设。例如当属性 E_l 为离散数据时，则假设 $P(E_l|\theta_j)$ 为多项式分布，而利用 θ_j 中属性 E_l 发生的比例得到 $P(E_l|\theta_j)$，若训练数据中类别为 θ_j 的数据笔数为 m_j，而所有满足 E_l 下且相依变量类别为 θ_j 的数据笔数为 r_{lj}，则 $P(E_l|\theta_j)$ 为

$$P(E_l \mid \theta_j)=\frac{r_{lj}}{m_j} \tag{7.11}$$

当属性 E_l 为连续数据时，则可利用训练数据配适连续型先验分布(例如高斯分布)求解。

［**范例 7.2**］　不动产公司搜集了 10 笔顾客数据，包括婚姻、年龄、收入等三个类别属性，目标变量为是否有购买不动产；假设 θ_1 表示购买了不动产，θ_2 代表没有购买不动产。一般而言，是否已经购买不动产的顾客其购买的动机会有所不同，因而影响销售人员的销售策略，若今天不动产经理人认识一位新的顾客，想根据其问卷所搜集的数据推测该顾客有无购买不动产，作为后续的销售规划依据。如顾客的属性数据 $\boldsymbol{E}^*=$(婚姻 $E_1^*=$ 已婚，年龄层 $E_2^*=$ 中年，收入 $E_3^*=$ 高)。

表 7.1 不动产公司顾客事务数据

ID	婚姻 E_1^*	年龄层 E_2^*	收入 E_3^*	购买不动产决策变数 $\tilde{\theta}$
001	已婚	青年	低	有
002	已婚	中年	高	无
003	单身	中年	高	无
004	单身	青年	高	有
005	已婚	中年	中	有
006	单身	中年	低	有
007	单身	青年	高	无
008	已婚	青年	高	无
009	已婚	中年	高	有
010	已婚	青年	高	有

如前所述，先验概率 $P(\tilde{\theta})$可由训练数据计算而得

$$P(\theta_1) = P(\text{购买了不动产}) = 6/10 = 0.60$$

$$P(\theta_2) = P(\text{没购买不动产}) = 4/10 = 0.40$$

因此当无任何其他信息时，可合理猜测来访的顾客，购买了不动产的概率为 0.6。

若加上属性的信息，可得 $P(\boldsymbol{E}^* | \theta_j)$的条件概率，以下先考虑仅由婚姻属性预测该顾客是否已经购买不动产：

$$\begin{aligned}&P(\text{购买不动产} = \text{有} \mid \text{婚姻} = \text{已婚})\\=&\frac{P(\text{已婚} \mid \text{购买了})P(\text{购买了})}{P(\text{已婚} \mid \text{购买了})P(\text{购买了}) + P(\text{已婚} \mid \text{没购买})P(\text{没购买})}\\=&\left(\frac{4}{6} \times \frac{6}{10}\right)\Big/\left(\frac{4}{6} \times \frac{6}{10} + \frac{2}{4} \times \frac{4}{10}\right) = 0.67\end{aligned}$$

$$\begin{aligned}&P(\text{购买不动产} = \text{无} \mid \text{婚姻} = \text{已婚})\\=&\frac{P(\text{已婚} \mid \text{没购买})P(\text{没购买})}{P(\text{已婚} \mid \text{购买了})P(\text{购买了}) + P(\text{已婚} \mid \text{没购买})P(\text{没购买})}\\=&\left(\frac{2}{4} \times \frac{4}{10}\right)\Big/\left(\frac{4}{6} \times \frac{6}{10} + \frac{2}{4} \times \frac{4}{10}\right) = 0.33\end{aligned}$$

发现考虑该顾客已经结婚下，推测该顾客可能已经购买不动产(0.67>0.33)，所以在销售规划可能就不适合以首次购屋的方式进行销售。

接着再考虑加入其他顾客的属性数据(婚姻、年龄层、收入)，再以朴素贝叶斯分类法计算其是否已经购买不动产，由表 7.1 可以直接估计 $P(\boldsymbol{E}^* | \theta_j)$：

$$P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_1) = \frac{P(\text{婚姻} = \text{已婚,年龄} = \text{中年,收入} = \text{高})}{P(\text{购买不动产} = \text{有})} = \frac{1/10}{6/10} = \frac{1}{6} \tag{7.12}$$

$$P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_2) = \frac{P(\text{婚姻} = \text{已婚,年龄} = \text{中年,收入} = \text{高})}{P(\text{购买不动产} = \text{无})} = \frac{1/10}{4/10} = \frac{1}{4} \tag{7.13}$$

由表 7.1 可以估计以下的概率：

$$P(E_1^* = \text{已婚} \mid \tilde{\theta} = \theta_1) = P(\text{婚姻} = \text{已婚} \mid \text{购买不动产} = \text{有}) = 4/6$$

$$P(E_1^* = \text{已婚} \mid \tilde{\theta} = \theta_2) = P(\text{婚姻} = \text{已婚} \mid \text{购买不动产} = \text{无}) = 2/4$$

$$P(E_2^* = \text{中年} \mid \tilde{\theta} = \theta_1) = P(\text{年龄层} = \text{中年} \mid \text{购买不动产} = \text{有}) = 3/6$$

$$P(E_2^* = \text{中年} \mid \tilde{\theta} = \theta_2) = P(\text{年龄层} = \text{中年} \mid \text{购买不动产} = \text{无}) = 2/4$$

$$P(E_3^* = \text{高} \mid \tilde{\theta} = \theta_1) = P(\text{收入} = \text{高} \mid \text{购买不动产} = \text{有}) = 3/6$$

$$P(E_3^* = \text{高} \mid \tilde{\theta} = \theta_2) = P(\text{收入} = \text{高} \mid \text{购买不动产} = \text{无}) = 4/4$$

若假设三个属性间为条件独立，根据以上的条件概率，可预测该顾客是否购买了不动产的计算结果如下：

$$\begin{aligned} &P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_1) = P(\boldsymbol{E}^* \mid \text{购买不动产} = \text{有}) \\ &\propto P(\text{婚姻} = \text{已婚} \mid \text{购买不动产} = \text{有}) \\ &\quad \cdot P(\text{年龄层} = \text{中年} \mid \text{购买不动产} = \text{有}) \\ &\quad \cdot P(\text{收入} = \text{高} \mid \text{购买不动产} = \text{有}) \\ &= \frac{4}{6} \times \frac{3}{6} \times \frac{3}{6} = \frac{1}{6} \end{aligned}$$

$$\begin{aligned} &P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_2) = P(\boldsymbol{E}^* \mid \text{购买不动产} = \text{无}) \\ &\propto P(\text{婚姻} = \text{已婚} \mid \text{购买不动产} = \text{无}) \\ &\quad \cdot P(\text{年龄层} = \text{中年} \mid \text{购买不动产} = \text{无}) \\ &\quad \cdot P(\text{收入} = \text{高} \mid \text{购买不动产} = \text{无}) \\ &= \frac{2}{4} \times \frac{2}{4} \times \frac{4}{4} = \frac{1}{4} \end{aligned}$$

上述计算结果，与式(7.12)及式(7.13)相同，可检验在给定有无购买不动产下，三个属性间为条件独立。可采用朴素贝叶斯分类法，再经由式(7.10)可得验后概率如下：

$$\begin{aligned} P(\tilde{\theta} = \theta_1 \mid \boldsymbol{E}^*) &= \frac{P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_1)P(\tilde{\theta} = \theta_1)}{P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_1)P(\tilde{\theta} = \theta_1) + P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_2)P(\tilde{\theta} = \theta_2)} \\ &= 0.167 \times 0.60/(0.167 \times 0.60 + 0.25 \times 0.40) = 0.5 \end{aligned}$$

$$\begin{aligned} P(\tilde{\theta} = \theta_2 \mid \boldsymbol{E}^*) &= \frac{P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_2)P(\tilde{\theta} = \theta_2)}{P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_1)P(\tilde{\theta} = \theta_1) + P(\boldsymbol{E}^* \mid \tilde{\theta} = \theta_2)P(\tilde{\theta} = \theta_2)} \\ &= 0.25 \times 0.40/(0.167 \times 0.60 + 0.25 \times 0.40) = 0.5 \end{aligned}$$

以该数据而言，表示仅依照加入年龄、收入属性后可能无法有效辨别该顾客是否已经购买不动产。在考虑三个属性后推测购买不动产的验后概率下降，可能是因为年龄与收入对于顾客是否购买不动产并非重要的属性，而属性是否重要则可经由主观经验或是利用属性筛选的方法，详细内容可见第 2 章。

[范例 7.2]的朴素贝叶斯分类法推导过程可用图 7.3 说明。图 7.3(a)为不具任何证据下推导可能的类别，主要是依据类别的历史数据。图 7.3(b)则为加入单一证据年龄下有无购买不动产的概率；图 7.3(c)则从婚姻、年龄、收入等三个证据推导有无购买不动产的概率；图 7.3(d)则将三个证据个别得到的概率相乘。从本范例中，有无考虑条件独立的概率计算结果相同，代表属性为条件独立。计算上也往往假设属性间为条件独立，因此应注意架

构不确定因子之间的关系时，是否符合条件独立的假设。

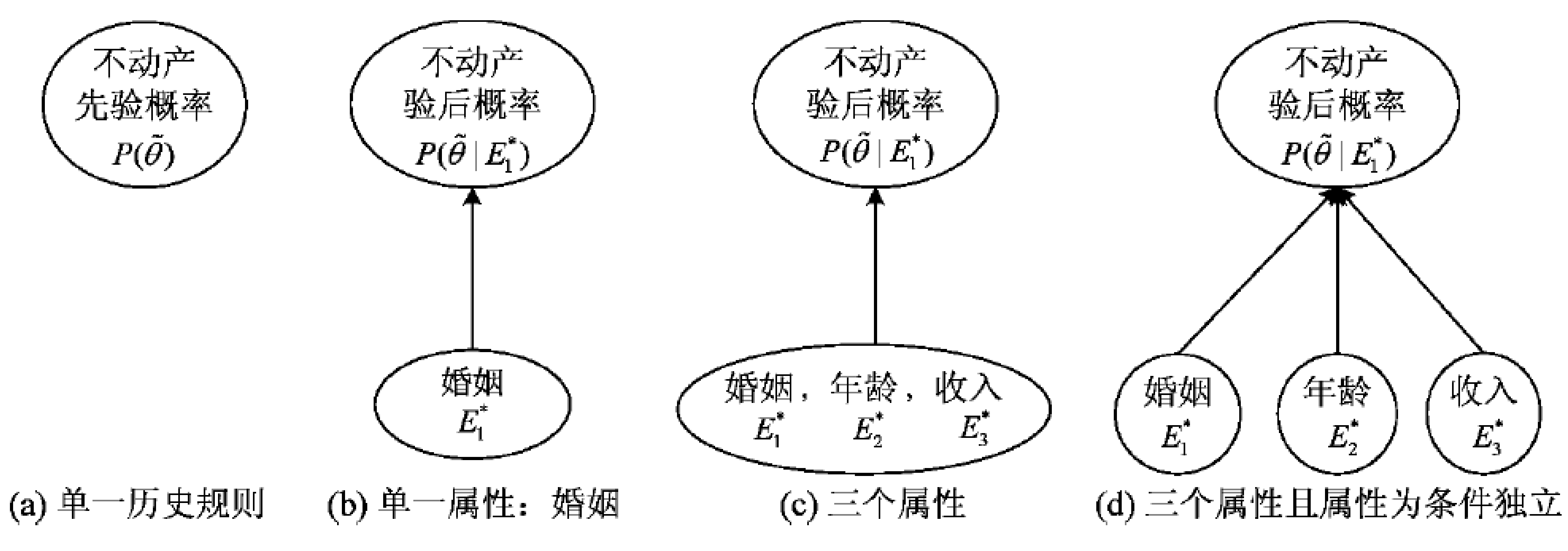


图 7.3 ［范例 7.2］推理关系

7.3 贝叶斯网络

朴素贝叶斯分类法假设属性之间对同一类别的影响互为条件独立，但实务上属性间往往存在相依关系，亦或者一个目标事件的推理通常需要多个证据。例如判断一位病患是否罹患癌症可能需要血液分析、尿液分析、超声波与触诊等结果汇整后，才能做判断。**贝叶斯网络（Bayesian networks）**是一种以图形呈现的统计推理（statistical inference）模型，将多个不确定事件利用一组随机变量以及变量间的影响关系来分析，并能随时根据新的信息或证据，通过层层推演，以修正相关的不确定事件的验后概率（Friedman *et al.*，1997）。

构建贝叶斯网络是将一个复杂且范围广泛的目标假设的不确定性判断，解析为多个有影响关系的不确定事件，每个不确定事件与目标假设的推论关系都是一个简单判断；并借由网络来表达简单节点之间的因果推论关系，经由分解再组合的过程，决策者可针对目标假设的评估，由最底层节点观察到的证据或样本信息，在网络架构中逐层推演更新而产生。

7.3.1 贝叶斯网络的理论基础

贝叶斯网络是用来处理复杂的推论关系，其中的每一个节点代表一个不确定事件，箭头代表推论法则的推论方向，以一箭头连接两节点表示一个法则。一个完整的贝叶斯推理网络除了网络图外，还需包含每一个节点的先验概率与每一个推论法则的强度（λ 与 $\bar{\lambda}$），也就是证据或样本信息的**似然函数（likelihood function）**或**似然比（likelihood ratio）**。

贝叶斯网络是“有向性的非循环图形”，亦即有关联的节点之间均以有方向性的箭头连结其推论关系，且不能有循环产生。贝叶斯网络节点间的连接关系依照证据与目标事件的推理关系可区分为：①单一证据（single evidence）推理关系，即只有一个证据节点指向一个目标事件节点；②多重证据推理关系（multiple evidence），即有多个证据节点指向一个目标事件节点；③多层次（multiple layer）推理关系，亦即经过两层以上的证据节点指向一个目标事件节点（简祯富，2014b）。图 7.4 为这三种推理关系的影响图，说明如下。

1. 单一证据推论

单一证据推论是统计推论的最基本形态。通常以 $\tilde{\theta}=H$ 表示一个决策者有兴趣的目

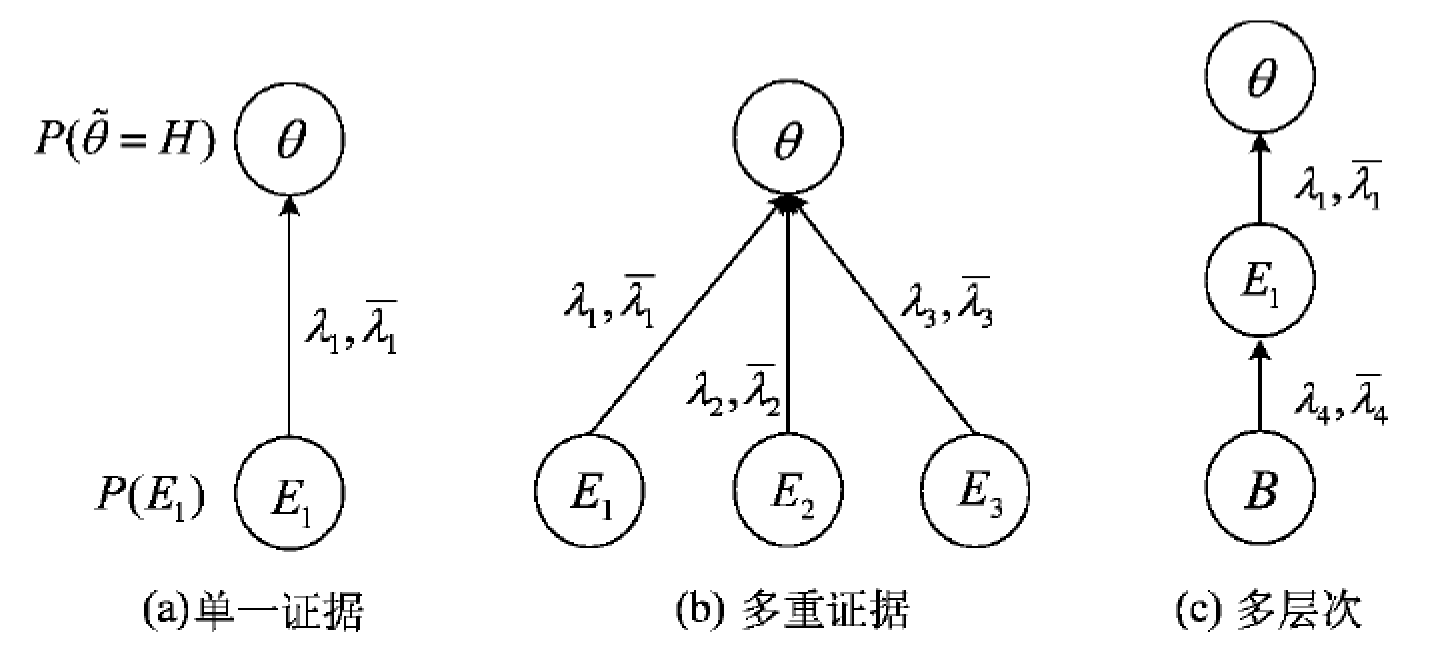


图 7.4　贝叶斯网络的三种基本推理关系（数据源：简祯富，2014b）

标假设，这个假设通常是一个不确定事件，以概率 $P(\tilde{\theta}=H)$ 来表示先验概率，以 E 表示一个有关的证据，单一证据推论关系可描述如下，并用式(7.14)表示：

$$\text{If}\quad E\quad \text{then}\quad \tilde{\theta}=H$$

$$P(H\mid E)=\frac{P(E\mid H)P(H)}{P(E\mid H)P(H)+P(E\mid \overline{H})P(\overline{H})} \tag{7.14}$$

在获得证据 E 后，不确定事件的验后概率可修正为 $P(\tilde{\theta}=H|E)$。例如一个参加定期健康检查的人想知道他是否有肝硬化的风险，H 代表肝硬化这个不确定事件，也就是目标假设，而 $P(\tilde{\theta})$ 则是肝硬化的先验概率。医生在未进行检查前，只能经由一般数据告诉他，有 1% 的国人会罹患肝硬化，即 $P(\tilde{\theta}=H)=1\%$。当医生发现该病患是 B 型肝炎带原者的新信息时，根据“若 B 肝带原，则罹患肝硬化”的推论，医生会修正他认为该病患罹患肝癌的概率为 P(肝硬化 | B 肝带原)＝25%。

似然函数 $P(E|H)$ 代表证据为 E 时，$\tilde{\theta}=H$ 的可能性，亦即证据 E 出现的概率随着给定不同的 $\tilde{\theta}$ 条件而变化。由于式(7.14)的分母为一个定值，因此也可表示为先验概率与验后概率的正比关系，如式(7.15)：

$$P(\tilde{\theta}\mid E)\propto P(\tilde{\theta})\cdot P(E\mid \tilde{\theta}) \tag{7.15}$$

先验概率可由三种方式取得(Berger, 1985)：①大量的先验信息，如历史数据，可利用数据分析或数据挖掘的方法计算概率；②含糊的先验知识，可由专家判断或决策者估计主观概率；③无先验数据提供任何信息，则可假设各种状态的概率相等。

以比率关系来表示 H 发生和 H 不发生的比率称为“**胜算**”(**odds**)，H 的先验胜算定义如式(7.16)：

$$O(H)=\frac{P(\tilde{\theta}=H)}{P(\tilde{\theta}=\overline{H})}=\frac{P(H)}{1-P(H)} \tag{7.16}$$

在确认 E 成立后，H 的验后胜算则可写为式(7.17)：

$$O(\tilde{\theta}=H\mid E)=\frac{P(\tilde{\theta}=H\mid E)}{P(\tilde{\theta}=\overline{H}\mid E)}=\frac{P(E\mid H)}{P(E\mid H)}\cdot\frac{P(H)}{P(H)} \tag{7.17}$$

同样地，也可以将 $\tilde{\theta}=H$ 的似然函数与 $\tilde{\theta}=\overline{H}$ 的似然函数以比率方式表达，称为似然比 λ，定义如式(7.18)：

$$\lambda = \frac{P(E \mid \tilde{\theta} = H)}{P(E \mid \tilde{\theta} = \overline{H})} \tag{7.18}$$

将似然比带入式(7.17)后改写为式(7.19)，亦即 H 的验后胜算等于 E 对 H 的似然比乘以 H 的先验胜算：

$$O(H \mid E) = \lambda \cdot O(H) \tag{7.19}$$

似然比 λ 可作为 E 确定成立时所提供的信息量指针，当 λ 越大时，表示 $\tilde{\theta}=H$ 时观察到 E 成立的可能性越高，且 $\tilde{\theta}=\overline{H}$ 时观察到 E 成立的可能性越低。换言之，决策者要推论 $\tilde{\theta}=H$ 是否为真时，似然比 λ 越高的证据 E 提供的参考信息越有说服力。

同理，若样本数据中显示证据 E 不存在($\overline{E}$)时，亦可根据似然比的定义推得 $\overline{E}$ 的似然比 $\overline{\lambda}$ 为

$$\overline{\lambda} = \frac{P(\overline{E} \mid \tilde{\theta} = H)}{P(\overline{E} \mid \tilde{\theta} = \overline{H})} = \frac{1 - P(E \mid H)}{1 - P(E \mid \overline{H})} \tag{7.20}$$

并根据 $\overline{E}$ 的似然比 $\overline{\lambda}$，建立当证据 E 不成立时 $\tilde{\theta}=H$ 之先验胜算与验后胜算的修正关系，如式(7.21)：

$$O(H \mid \overline{E}) = \overline{\lambda} \cdot O(H) \tag{7.21}$$

事实上，先验概率 $P(H)$ 与胜算 $O(H)$ 对 $\tilde{\theta}=H$ 的假设提供相同的信息，二者的转换关系式如式(7.22)：

$$P(H) = \frac{O(H)}{1 + O(H)} \tag{7.22}$$

同样地，验后概率 $P(H|E)$ 与验后胜算 $O(H|E)$ 的关系为

$$P(H \mid E) = \frac{O(H \mid E)}{1 + O(H \mid E)} \tag{7.23}$$

而 λ 和 $\overline{\lambda}$ 二者之间的关系并非完全独立，$\lambda>1$ 则 $\overline{\lambda}\leqslant 1$，$\overline{\lambda}>1$ 则 $\lambda\leqslant 1$，但两者亦非简单的互补关系，$\overline{\lambda}$ 可根据 λ 推导出两者的关系，如式(7.24)：

$$\overline{\lambda} = \frac{1 - P(E \mid H)}{1 - P(E \mid \overline{H})} = \frac{1 - \lambda \cdot P(E \mid \overline{H})}{1 - P(E \mid \overline{H})} \tag{7.24}$$

一般而言，$\lambda>1$，则 $\overline{\lambda}\leqslant 1$，表示 H 发生的可能性随证据 E 成立而增加；$\overline{\lambda}>1$，则 $\lambda\leqslant 1$，表示 H 发生的可能性随证据 E 不成立而增加。λ 或 $\overline{\lambda}$ 等于 1 分别表示 H 无法根据证据 E 成立或不成立进一步判断，亦即 H 的概率仍保持其先验概率。

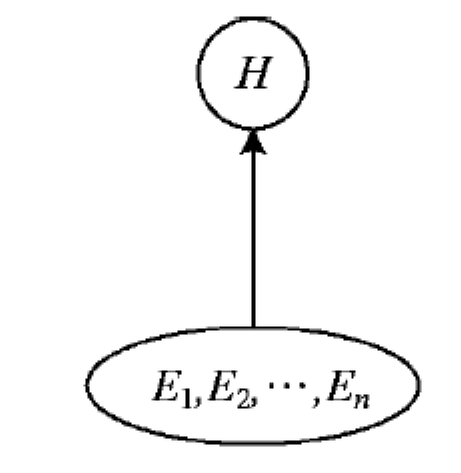
2. 多重证据推论

如果推论时参考的信息或观察的证据不只一种，则为多重证据推论，其影响图如图 7.5。贝叶斯网络的多重证据推论规则为

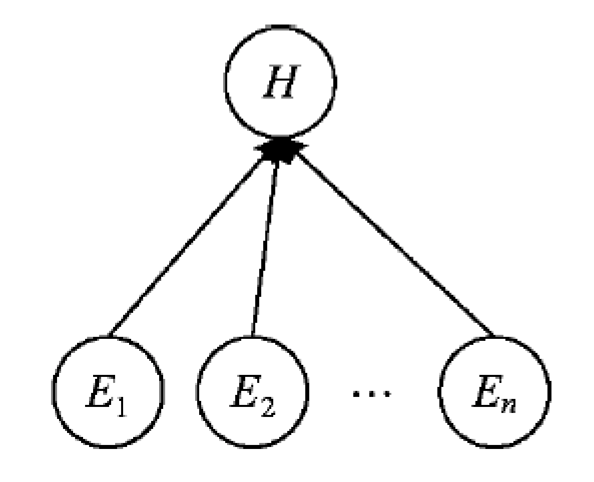
$$\text{If} \quad E_1 \quad \text{and} \quad E_2 \quad \text{and} \quad \cdots \quad \text{and} \quad E_n, \quad \text{then} \quad \tilde{\theta} = H$$

其中，E_i 表示第 i 个证据，则证据 $E_1,E_2,\cdots,E_n$ 成立时，H 的验后概率可写为

$$P(H \mid E_1, E_2, \cdots, E_n) = \frac{P(E_1, E_2, \cdots, E_n \mid H) \cdot P(H)}{P(E_1, E_2, \cdots, E_n)} \tag{7.25}$$




(a) 证据$E_1, E_2, \cdots, E_n$联合



(b) 证据$E_1, E_2, \cdots, E_n$为条件独立

图 7.5　多重证据推论示意图

根据贝叶斯定理,若证据 $E_1, E_2, \cdots, E_n$ 在给定 $\tilde{\theta}=H$ 时为条件独立,亦即每个证据 E_i 对 $\tilde{\theta}=H$ 的可能性或似然函数 $P(E_i|\tilde{\theta}=H)=\lambda_i$,均不受其他证据和 $\tilde{\theta}=H$ 的推理关系影响,如图 7.5(b)所示,则 $E_1, E_2, \cdots, E_n$ 对 $\tilde{\theta}=H$ 的联合似然函数 $P(E_1, E_2, \cdots, E_n|\tilde{\theta}=H)$为个别似然函数 $P(E_i|\tilde{\theta}=H)$的乘积:

$$P(E_1, E_2, \cdots, E_n \mid \tilde{\theta} = H) = \prod_{i=1}^{n} P(E_i \mid \tilde{\theta} = H)$$

因此式(7.25)可改写为式(7.26):

$$P(\tilde{\theta} = H \mid E_1, E_2, \cdots, E_n) = \frac{\prod_{i=1}^{n} P(E_i \mid \tilde{\theta} = H) \cdot P(\tilde{\theta} = H)}{P(E_1, E_2, \cdots, E_n)} \tag{7.26}$$

同理,若 $E_1, E_2, \cdots, E_n$ 在给定 $\tilde{\theta}=\overline{H}$ 时为条件独立,则 $\tilde{\theta}=\overline{H}$ 的验后概率为式(7.27):

$$P(\tilde{\theta} = \overline{H} \mid E_1, E_2, \cdots, E_n) = \frac{\prod_{i=1}^{n} P(E_i \mid \tilde{\theta} = \overline{H}) \cdot P(\tilde{\theta} = \overline{H})}{P(E_1, E_2, \cdots, E_n)} \tag{7.27}$$

将式(7.26)与式(7.27)相除可得多重证据推论时的 $\tilde{\theta}=H$ 验后胜算:

$$O(\tilde{\theta} = H \mid E_1, E_2, \cdots, E_n) = O(\tilde{\theta} = H) \cdot \prod_{i=1}^{n} \lambda_i \tag{7.28}$$

其中,λ_i 为证据 E_i 成立的似然比,$\lambda_i = \dfrac{P(E_i|\tilde{\theta}=H)}{P(E_i|\tilde{\theta}=\overline{H})}$。

同理,若证据 E_i 不成立,以 $\bar{E}_i$ 表示之,也就是 $E_1, E_2, \cdots, E_n$ 成立,且 $E_1, E_2, \cdots, E_n$ 对 $\tilde{\theta}=H$ 和 $\tilde{\theta}=\overline{H}$ 皆为条件独立,则 $\tilde{\theta}=H$ 的验后胜算为

$$O(\tilde{\theta} = H \mid \bar{E}_1, \bar{E}_2, \cdots, \bar{E}_n) = \left(\prod_{i=1}^{n} \bar{\lambda}_i\right) \cdot O(\tilde{\theta} = H) \tag{7.29}$$

其中,$\bar{\lambda}_i$ 为证据 E_i 不成立的似然比,$\bar{\lambda}_i = \dfrac{P(\bar{E}_i|\tilde{\theta}=H)}{P(\bar{E}_i|\tilde{\theta}=\overline{H})}$。

因此,在单一证据与多重证据的贝叶斯网络推论中,每一个推论关系都具有证据成立的似然比 λ_i 与证据不成立的似然比 $\bar{\lambda}_i$,分别代表 E_i 成立或不成立时对假设 $\tilde{\theta}=H$ 的修正及其强度。例如,当 $\lambda_i>1$ 表示证据 E_i 强化 $\tilde{\theta}=H$ 的胜算,反之,$\lambda_i<1$ 表示证据 E_i 弱化 $\tilde{\theta}=$

H 的胜算。由于 λ_i 与 $\bar{\lambda}_i$ 两者互有关联，因此由 $\lambda_i>1$ 即可推知 $\bar{\lambda}_i<1$，反之亦然。$\lambda_i=1$ 当且仅当 $\bar{\lambda}_i=1$，则代表 $\tilde{\theta}=H$ 为真与否完全不能由证据 E_i 判断，换言之，是否观察到 E_i 并没有改变对于 $\tilde{\theta}=H$ 发生概率的估计。

实务上，所有证据都符合条件独立的情况并不一定会成立，这时，不应为了使用贝叶斯网络的多重推论，而强行将证据 $E_1,E_2,\cdots,E_n$ 拆解成 n 个条件独立的证据。应当回归证据间的实际关系，将确实符合条件独立的证据区隔出来，而不符合条件独立的证据则保持相依的关系，如图 7.6 所示。进行贝叶斯推论时，亦仅区分出独立证据的似然概率，相依的证据则沿用联合似然函数，假定仅有证据 E_1 与 E_2 符合条件独立关系时，则

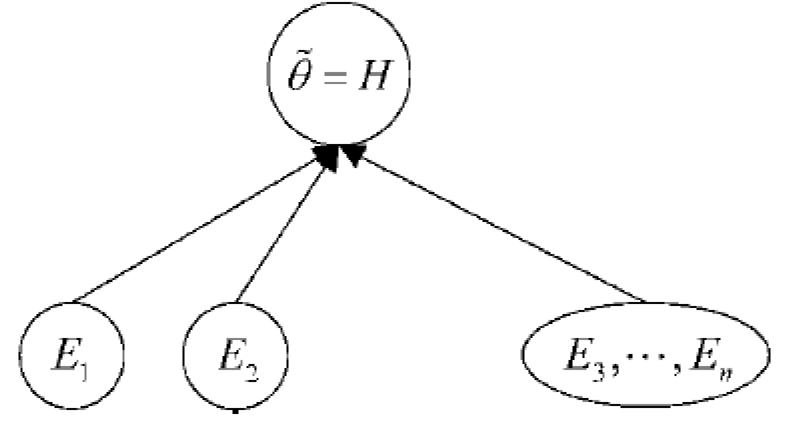


图 7.6 多重证据推论，只有部分证据满足条件独立的假设

$$P(E_1,E_2,\cdots,E_n \mid \tilde{\theta}=H)=P(E_1 \mid \tilde{\theta}=H)\cdot P(E_2 \mid \tilde{\theta}=H)$$
$$\cdot P(E_3,E_4,\cdots,E_n \mid \tilde{\theta}=H)$$

3. 多层推论

在多层次的贝叶斯网络中，节点间的因果关系较为复杂，网络中的某一节点，可能同时是其后续节点的因，也是前行节点的果。例如图 7.7(c)的节点 E 是节点 H 的因，因此 E 可作为推论 H 是否为真的证据；但 E 本身又是节点 B 的果，因此 E 是否为真，也是 B 作为证据要推论的假设。但每个节点都只和与其有方向性的箭头直接相连的节点有推论上的因果关系，贝叶斯网络的节点不会有循环产生(Chien，2005)。

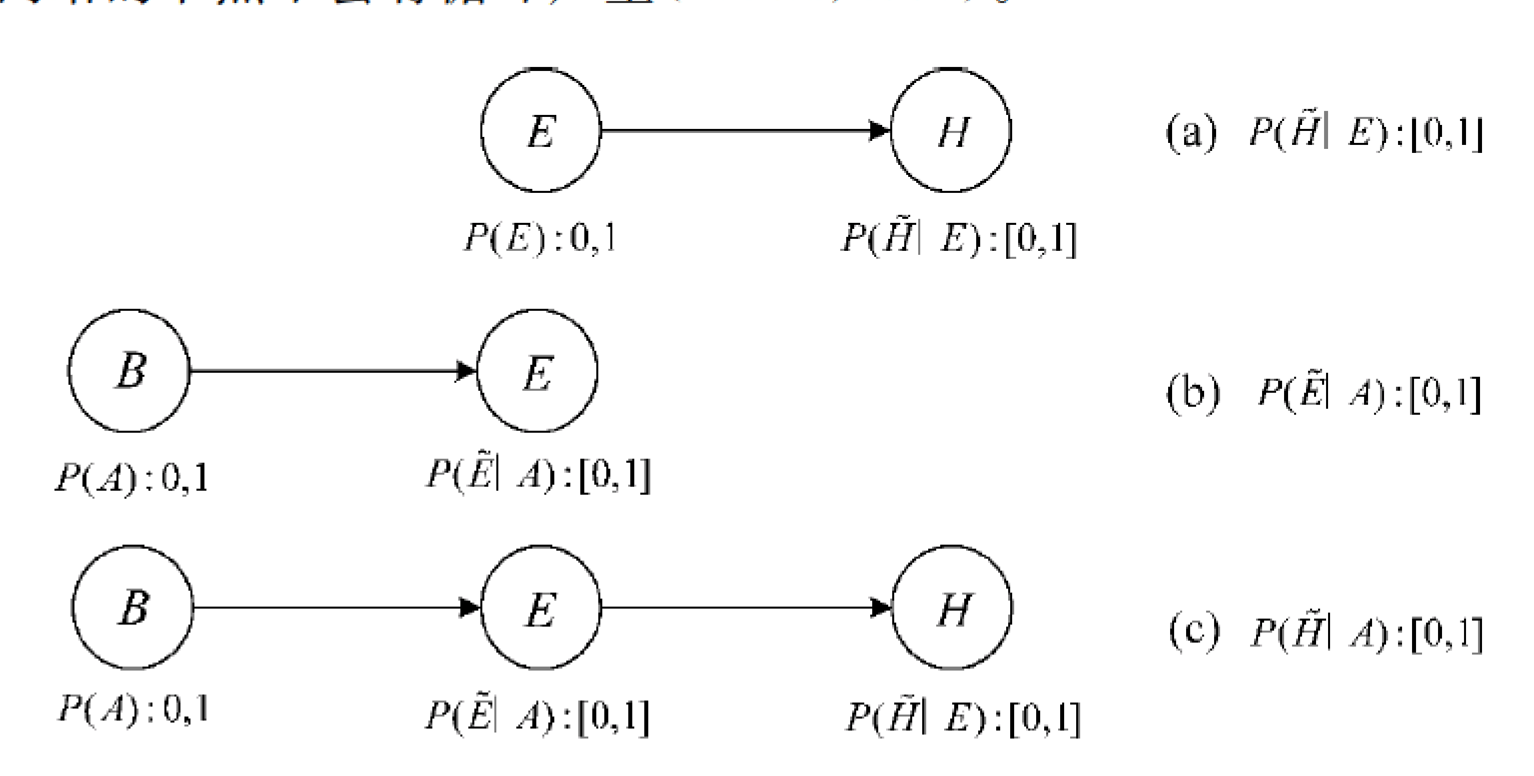


图 7.7 多层推论贝叶斯网络节点关系

当用户提供网络根部节点的任何新证据时，贝叶斯推理即往上逐层修正每一个节点的概率，直到最顶层的目标事件为止，如此即可求得每个事件的验后概率。

若考虑 E 的不确定性，亦即不确定证据 E 是否成立，仅能得知 E 成立的概率时，则须适当修正上述贝叶斯定理的计算方式。假设经过观测事件 B 后，仅能在某些程度上确认 E 是否成立，则将 E 成立的概率表示为 $P(E|B)$。根据概率理论，可将 $P(H|B)$作适当转换：

$$P(H \mid B)=P(H,E \mid B)+P(H,\bar{E} \mid B)$$

$$=P(H \mid E,B)P(E \mid B)+P(H \mid \bar{E},B)P(\bar{E} \mid B) \tag{7.30}$$

当证据 E 已确定是否成立时，任何观测行为 B 都是多余的，因此：

$$P(H \mid E,B) = P(H \mid E)$$
$$P(H \mid \bar{E},B) = P(H \mid \bar{E})$$

根据上述条件，式(7.30)可简化为

$$P(H \mid B) = P(H \mid E)P(E \mid B)+P(H \mid \bar{E})P(\bar{E} \mid B) \tag{7.31}$$

证据 B 直接对 H 的有效似然比 λ_B 为

$$\lambda_B = \frac{P(B \mid H)}{P(B \mid H)}$$

则 λ_B 可改写为

$$\lambda_B = \frac{P(B \mid H)}{P(B \mid \bar{H})} = \frac{P(H \mid B)}{P(\bar{H} \mid B)} \cdot \frac{P(\bar{H})}{P(H)} = \frac{O(H \mid B)}{O(H)} \tag{7.32}$$

其中，$O(H|B)=\dfrac{P(H|B)}{P(\bar{H}|B)}=\dfrac{P(H|B)}{1-P(H|B)}$。

7.3.2　贝叶斯网络的不一致性修正

由于在贝叶斯网络中，多层推论时的中间层节点是由其他机会节点推论而来，因此其状态不确定，故某节点的先验概率和由该节点的先行节点所推得的概率可能会产生不一致(inconsistent)。

由式(7.31)可知，尽管 E 是由 B 推论而来而具有不确定性，但在推论时，若观察到 B 可以完全确定 E 发生，也就是 $P(E|B)=1$ 而 $P(\bar{E}|B)=0$，或是完全确定 E 不发生，也就是 $P(\bar{E}|B)=1$ 而 $P(E|B)=0$，这两个特例就如同图 7.8(a)中的两个端点。由式(7.32)可知，$P(E|B)$对 $P(H|B)$的图形为由上述两个端点构成的线段，如图 7.8(b)所示，因此当 $P(E|B)$已知时，根据图 7.8(c)中的线段即可推得对应的 $P(H|B)$。

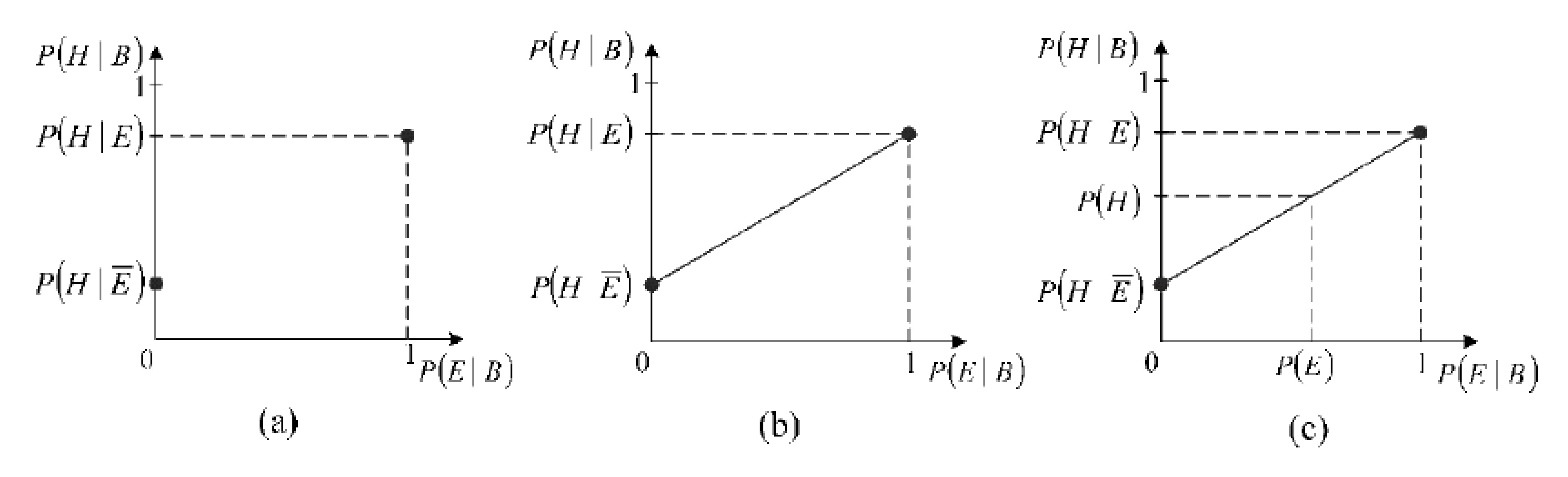


图 7.8　理想状况下 $P(E|B)$对 $P(H|B)$的关系图

理论上，当 $P(E|B)=P(E)$时，表示 B 的观测对于判断证据 E 成立与否并无贡献，根据式(7.31)可得出 $P(H|B)=P(H)$，也就是说 H 的验后概率与先验概率相同，即 $P(E)$和 $P(H)$对应的点应落于图 7.8(b)线段上，如图 7.8(c)所示。然而，在实务上可能产生 $P(E|B)$等于 $P(E)$时，$P(H|B)$却不等于 $P(H)$的不一致现象，如图 7.9(a)与图 7.9(b)所示。

学者(Duda *et al.*，1979，1976)提出几种不同的修正方法，详细的比较和讨论可参考(Chien，2005)。其中，线性内插函数(linear interpolation functions)方法，将不一致的问题修正如式(7.33)，经线性内插修正后，$P(E|B)$对 $P(H|B)$的关系图则如图 7.10

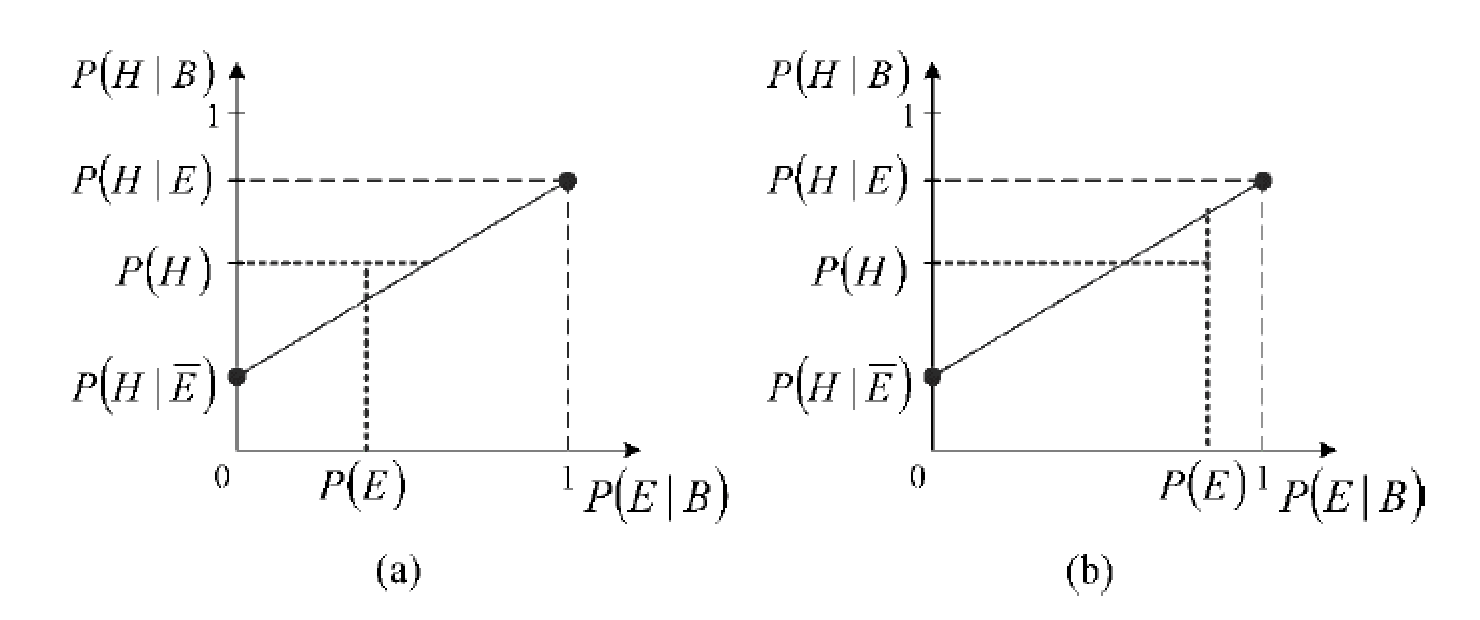


图 7.9 不一致状况下 $P(E|B)$对 $P(H|B)$的关系图

所示。

$$P(H \mid B)=\begin{cases}P(H)+\dfrac{P(E \mid B)-P(E)}{1-P(E)} \cdot [P(H \mid E)-P(H)], & P(E \mid B) \geqslant P(E) \\ P(H)-\dfrac{P(E \mid B)-P(E)}{P(E)} \cdot [P(H \mid \bar{E})-P(H)], & P(E \mid B)<P(E)\end{cases} \tag{7.33}$$

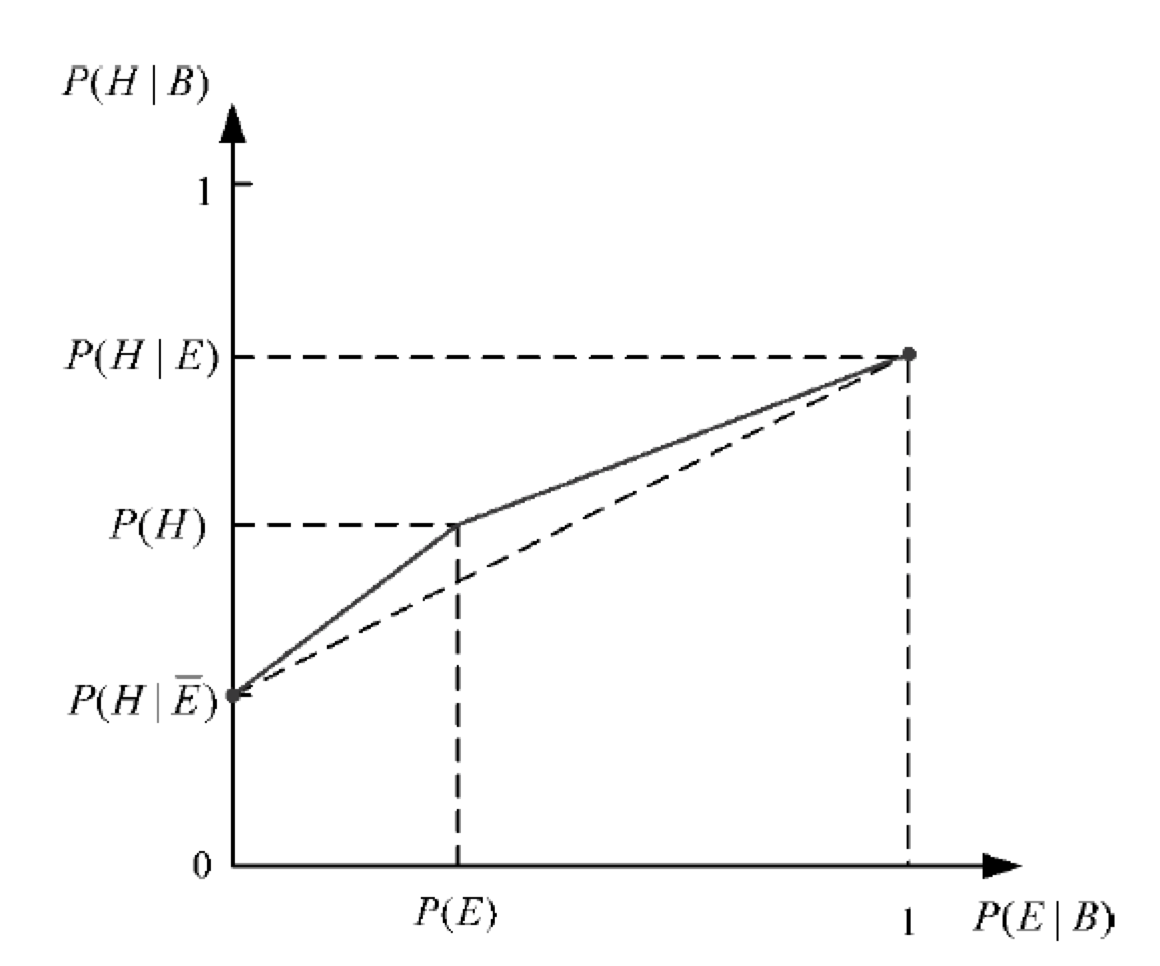


图 7.10 线性内插修正后 $P(E|B)$对 $P(H|B)$的关系图

将许多层级与证据构成如图 7.11 所示的贝叶斯推理网络，每一个节点必须储存该事件的先验概率，每一个箭头须储存该推论法则的强度 λ 与 $\bar{\lambda}$。当用户提供网络底层的任何证据时，即根据式(7.32)以及贝叶斯网络的推论路径，计算前面节点的证据对目标事件的有效似然比，并配合式(7.33)的修正式，逐层修正每一个节点的概率，直到目标节点为止，如此即可求得每个事件的验后概率。

总而言之，贝叶斯网络包含一组以单一证据、多重证据与多层次的推论关系所连接的节点，将复杂的不确定事件分解并简化其推论关系后，再整合起来作综合推论。虽然贝叶斯网络的功能强大，然而相对于其他的机器学习方法，极耗计算时间。随着信息科技的提升，贝叶斯网络日益重要，例如搜索引擎 Google 与网络书店 Amazon 皆广泛使用此种方法。

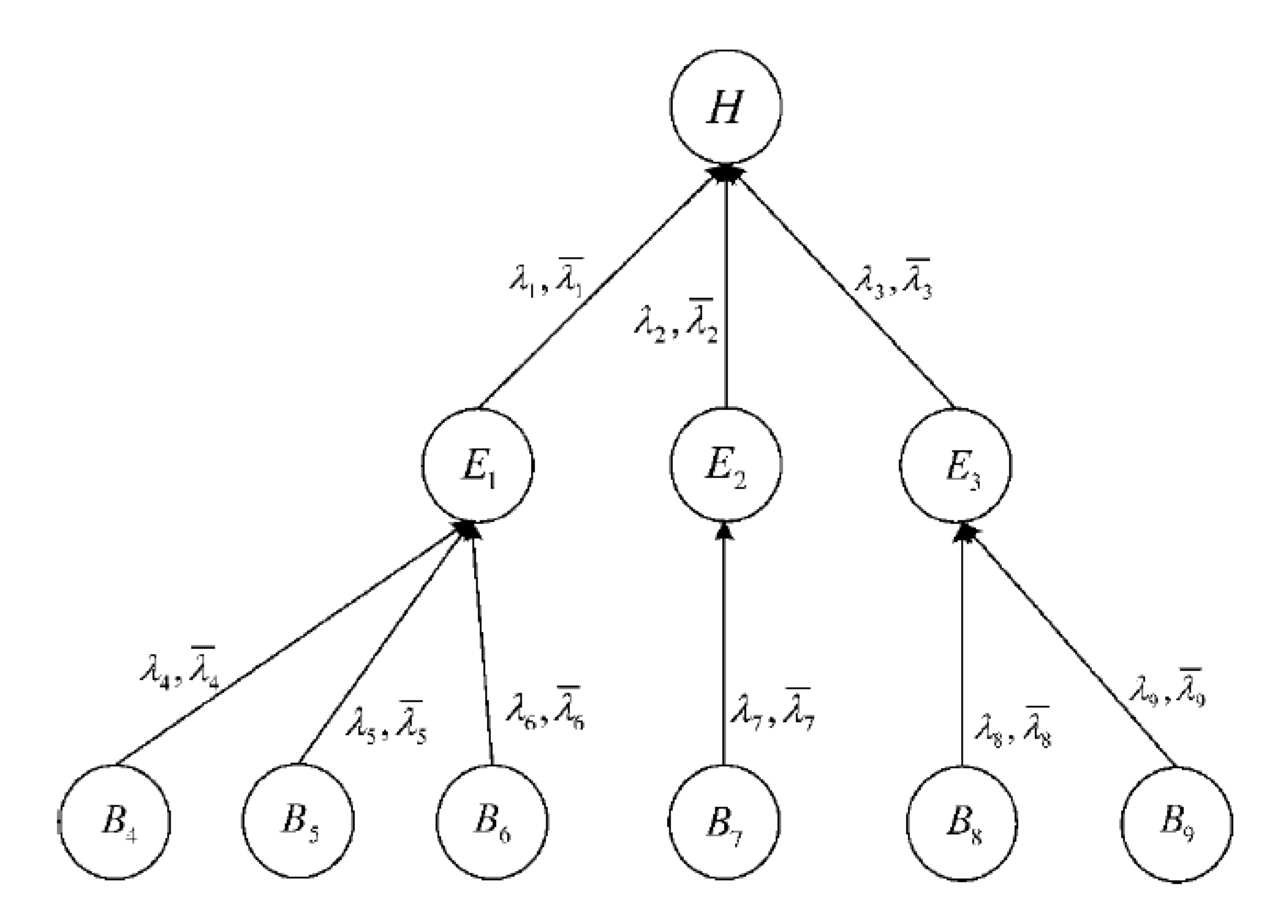


图 7.11　多层次、多重证据的贝叶斯网络图(数据源：简祯富,2014b)

7.4　R 语言与贝叶斯分类

本节将说明如何通过 R 语言应用朴素贝叶斯分类法与贝叶斯网络,两种方法都内建在扩充套件 **bnlearn**(Scutari,2014)。

延续皮马族印第安人糖尿病数据集,借由怀孕次数(npreg)、葡萄糖浓度(glu)、血压(bp)、三头肌皮褶厚度(skin)、身体质量指数(bmi)、糖尿病家族病因指数(ped)与年龄(age)等 7 个属性进行验后概率的推论,以判断是否罹患糖尿病(type)。首先,由于贝叶斯分类仅支持类别型变量,因此需将 7 个连续型属性进行离散化。在此,将所有数据合并后对每一个属性利用等宽分箱法进行 2 等份的离散化,再将前 200 笔切割为训练集数据、后 332 笔为测试集数据。离散化后的各属性水平区间定义如表 7.2 所示。

表 7.2　离散化后属性水平区间

属 性	水平 1	水平 2	属 性	水平 1	水平 2
npreg	[−0.017,8.5]	(8.5,17]	bmi	[18.2,42.6]	(42.6,67.1]
glu	[55.9,128]	(128,199]	ped	[0.0827,1.25]	(1.25,2.42]
bp	[23.9,67]	(67,110]	age	[20.9,51]	(51,81.1]
skin	[6.91,53]	(53,99.1]			

```
library(MASS)
library(RSNNS)
data("Pima.tr")
data("Pima.te")
set.seed(1111)
Pima= rbind(Pima.tr,Pima.te)
level_name= {}
for (i in 1:7) {
```

```
  Pima[,i]=cut(Pima[,i],breaks=2,ordered_result=T,include.lowest=T)
  level_name<-rbind(level_name,levels(Pima[,i]))
}
level_name=data.frame(level_name)
row.names(level_name)=colnames(Pima)[1:7]
colnames(level_name)=paste("L",1:2,sep="")
level_name
Pima.tr=Pima[1:200,]
Pima.te=Pima[201:nrow(Pima),]
```

接着通过扩充套件 **bnlearn** 中 **naive. bayes** 函数建立朴素贝叶斯分类法，并建立其网络结构图，如图 7.12 中可发现朴素贝叶斯分类法是架构在 7 属性是互为独立且均只单独受到 type 影响的假设下运行，因此模型中需要进行估计的先验概率，包含 $P(\text{type})$、$P(\text{npreg}|\text{type})$、$P(\text{glu}|\text{type})$、$P(\text{bp}|\text{type})$、$P(\text{skin}|\text{type})$、$P(\text{bmi}|\text{type})$、$P(\text{ped}|\text{type})$、$P(\text{age}|\text{type})$ 等 8 项，而对于每一笔的观察值而言其属性发生的联合概率为个别概率的相乘值：

$$
\begin{aligned}
P(\text{npreg},\text{glu},\text{bp},\text{skin},\text{bmi},\text{ped},\text{age} \mid \text{type}) = & P(\text{npreg} \mid \text{type}) \cdot P(\text{glu} \mid \text{type}) \\
& \cdot P(\text{bp} \mid \text{type}) \cdot P(\text{skin} \mid \text{type}) \\
& \cdot P(\text{bmi} \mid \text{type}) \cdot P(\text{ped} \mid \text{type}) \\
& \cdot P(\text{age} \mid \text{type})
\end{aligned}
$$

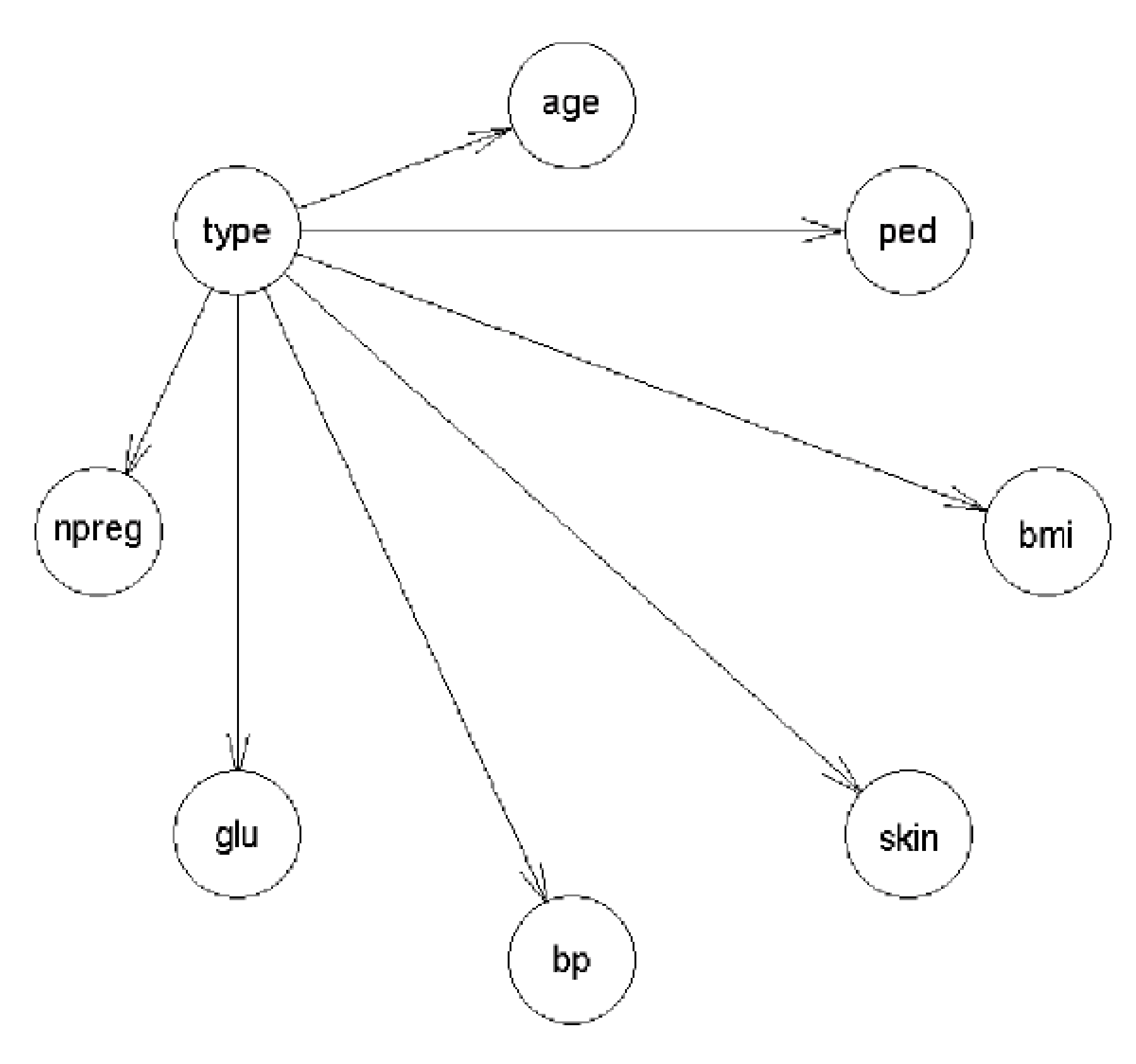


图 7.12 罹患糖尿病的朴素贝叶斯分类法网络结构图

表 7.3 说明 7 个属性受 type 影响的先验概率估计值，经由贝叶斯定理可推导给定属性值之下的验后概率，作为推论预测测试数据中的 type 属性。在 332 笔的测试数据中，经由朴素贝叶斯分类法的结果正确的有 250 笔，正确率为 75.3%。

```
library(bnlearn)
bn=naive.bayes(Pima.tr,"type");plot(bn);bn
fitted=bn.fit(bn,Pima.tr)
pred=predict(fitted,Pima.te)
```

```
tab=table(pred,Pima.te[,"type"]);tab
acc=sum(diag(tab))/sum(tab);acc
```

表 7.3　朴素贝叶斯分类法先验概率估计值

属　　性	水　平	type=No	type=Yes
怀孕次数	[−0.017,8.5]	0.939	0.824
	(8.5,17]	0.061	0.176
葡萄糖浓度	[55.9,128]	0.758	0.279
	(128,199]	0.242	0.721
血压	[23.9,67]	0.402	0.250
	(67,110]	0.598	0.750
三头肌皮褶厚度	[6.91,53]	0.992	0.985
	(53,99.1]	0.008	0.015
身体质量指数	[18.2,42.6]	0.955	0.941
	(42.6,67.1]	0.045	0.059
糖尿病家族病因指数	[0.0827,1.25]	0.992	0.956
	(1.25,2.42]	0.008	0.044
年龄	[20.9,51]	0.955	0.838
	(51,81.1]	0.045	0.162

接着通过 **tree. bayes** 函数构建贝叶斯网络，其网络结构图如图 7.13 所示，用户可依据自身对于属性间因果关联性，通过 whitelist 与 blacklist 等函数自变量自行添加或取消连结箭头，并进行先验概率的估计。

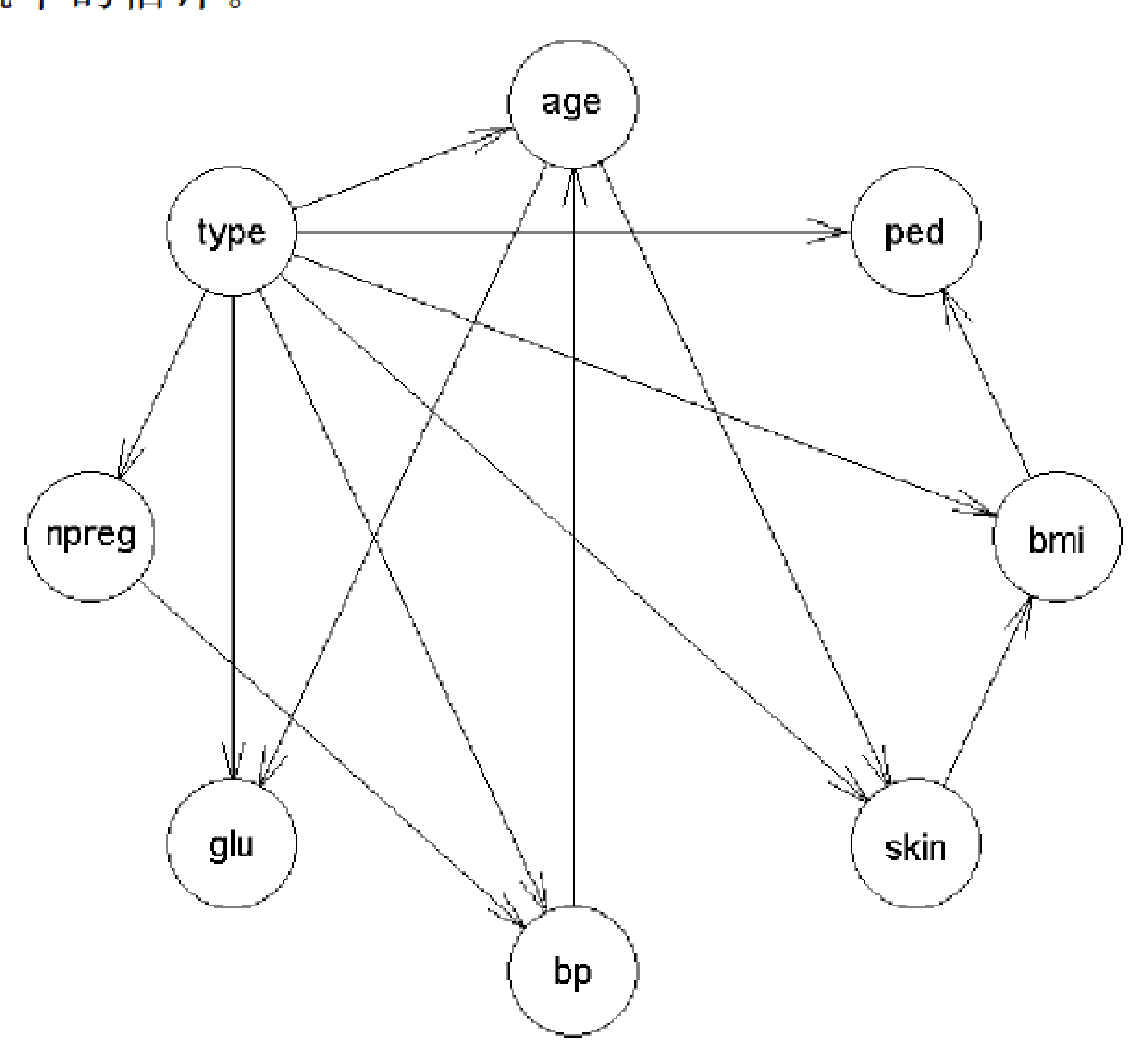


图 7.13　是否罹患糖尿病的贝叶斯网络结构图

```
tan=tree.bayes(Pima.tr,"type");plot(tan);tan
#whitelist 自变量可设定要增加的连结箭头;blacklist 自变量可设定要取消的箭头
fitted=bn.fit(tan,Pima.tr,method="bayes")
pred=predict(fitted,Pima.te)
tab=table(pred,Pima.te[,"type"]);tab
acc=sum(diag(tab))/sum(tab);acc
```

在此例中,共需要估计 8 项先验概率,包含 P(type)、P(npreg|type)、P(bp|type,npreg)、P(age|type,bp)、P(glu|type,age)、P(skin|type,age)、P(bmi|type,skin)、P(ped|type,bmi)。相较于朴素贝叶斯分类法,除了前两项相同之外,后面 6 个属性的先验概率并不单只受 type 影响,其先验概率值估计分别如表 7.4 至表 7.9 所示。可推导出给定属性值之下的验后概率,用以预测测试数据中的 type 属性。在 332 笔的测试数据中,简单网络结果被正确分类的有 251 笔,正确率为 75.6%,与朴素贝叶斯分类法所得的结果差异不大。

表 7.4　贝叶斯网络分类器 glu 属性的先验概率估计值

		glu=[55.9,128]	glu=(128,199]	合　计
age=[20.9,51]	type=No	0.772	0.228	1
	type=Yes	0.307	0.693	1
age=(51,81.1]	type=No	0.382	0.618	1
	type=Yes	0.241	0.759	1

表 7.5　贝叶斯网络分类器 bp 属性的先验概率估计值

		bp=[23.9,67]	bp=(67,110]	合　计
npreg=[−0.017,8.5]	type=No	0.429	0.571	1
	type=Yes	0.278	0.722	1
npreg=(8.5,17]	type=No	0.119	0.881	1
	type=Yes	0.224	0.776	1

表 7.6　贝叶斯网络分类器 skin 属性的先验概率估计值

		skin=[6.91,53]	skin=(53,99.1]	合　计
age=[20.9,51]	type=No	0.982	0.018	1
	type=Yes	0.979	0.021	1
age=(51,81.1]	type=No	0.853	0.147	1
	type=Yes	0.833	0.167	1

表 7.7　贝叶斯网络分类器 age 属性的先验概率估计值

		age=[20.9,51]	age=(51,81.1]	合　计
bp=[23.9,67]	type=No	0.977	0.023	1
	type=Yes	0.885	0.115	1

续表

		age=[20.9,51]	age=(51,81.1]	合　计
bp=(67,110]	type=No	0.911	0.089	1
	type=Yes	0.790	0.210	1

表7.8　贝叶斯网络分类器bmi属性的先验概率估计值

		bmi=[18.2,42.6]	glu=(42.6,67.1]	合 计
skin=[6.91,53]	type=No	0.953	0.047	1
	type=Yes	0.924	0.076	1
skin=(53,99.1]	type=No	0.357	0.643	1
	type=Yes	0.643	0.357	1

表7.9　贝叶斯网络分类器ped属性的先验概率估计值

		ped=[0.0827,1.25]	glu=(1.25,2.42]	合 计
bmi=[18.2,42.6]	type=No	0.982	0.018	1
	type=Yes	0.966	0.034	1
bmi=(42.6,67.1]	type=No	0.853	0.147	1
	type=Yes	0.500	0.500	1

7.5　应用实例——电力公司馈线事故定位系统

7.5.1　案例简介与问题架构

配电高压馈线(feeder)事故定位是在馈线发生事故后,迅速检出故障区以加速隔离并转供其他电源,以缩短用户的停电时间,减少经济损失和社会成本。

在电力公司配电馈线尚未自动化前,当配电馈线发生故障时,馈线断路器迅速跳脱,导致全线用户停电。工作人员必须根据经验判断并立即赶赴事故现场,执行试送电,以确定故障地址,且试送电的动作不利于电缆绝缘,因此故障的检测相当耗费时间。即使已自动化的馈线,对于分歧在线的事故定位仍须依赖经验。

本案例(Chien, *et al.*, 2002)系针对某电力公司馈线事故维修数据,以贝叶斯网络发展数据挖掘方法作为发展事故定位专家系统之基础,实证研究结果验证本方法可以推论在不同的事故状况下各设备的相对损坏可能性。

贝叶斯网络的构建是一个反复构建、验证与修正的演进过程,需要通过领域知识提取的过程,并辅以统计数据的分析结果,以构建贝叶斯推理网络,共包含三个阶段:

(1) 确定目标假设,也就是最顶层的节点,以及与目标假设有关的随机变量,以确认假设的证据或观察数据。

(2) 建立变量之间的影响图,以定义变量间的相依性。其中,指向同一个节点的所有先

行节点必须为条件独立。

(3) 对每一个变量建立局部条件概率分布模式评估与分析。

7.5.2　数据整理与贝叶斯网络图构建

本案例与配电调度领域的专家进行多次结构性访谈(structured interview),来验证研究小组构建的贝叶斯推理网络模型与专家的推理逻辑是否一致。研究架构中一个主轴是分析专家的心智架构以建立贝叶斯网络模型,亦即建立假设与证据间的推理关系。另一个主轴则是撷取专家知识,对贝叶斯网络中每一个推理关系给定参数值作为输入项,包含 $P(H)$、$P(E_i)$、λ_i 与 $\bar{\lambda}_i$,将两个主轴结合即可得完整的贝叶斯网络并以实际数据进行验证。

首先分析某电力公司现行的配电系统事故停电统计数据,协助决定贝叶斯推理网络所需要的变量(节点)。主要的数据源是配电事故停电记录表,记录表中包含事故日期、时间、地点、发生事故的设备(例如变压器)、事故原因(例如火灾)等。为降低贝叶斯网络的复杂度以减少不必要的计算过程,研究小组分析频率、相关性、先验概率与条件概率,将相关的项目整合为单一项目以降低网络的节点数。另外,也多次与专家进行讨论,以求更清楚且实务地解读历史数据,并提升研究小组对电力配送系统的知识。

经由历史数据分析与领域知识整理后,构建出事故定位的贝叶斯网络初始模型,初始模型中的推理逻辑可分为四层,依序为:可观察的现象→事故原因→事故情形→损坏的设备,如图 7.14 所示。由于可能损坏的设备种类繁多,可观察的现象也很多,因此根据推理逻辑构建出的贝叶斯推理网络图非常复杂,图 7.15 为馈线事故定位推理网络图的一部分,包含油压转换器与地下转换器这两种设备损坏的推理网络图。贝叶斯推理网络的最下层为变电所可观察到的现象,即为事故定位专家系统的输入事实,最上层则为损坏的设备,即事故定位专家系统的输出结果,据以指出馈线最可能故障的设备。

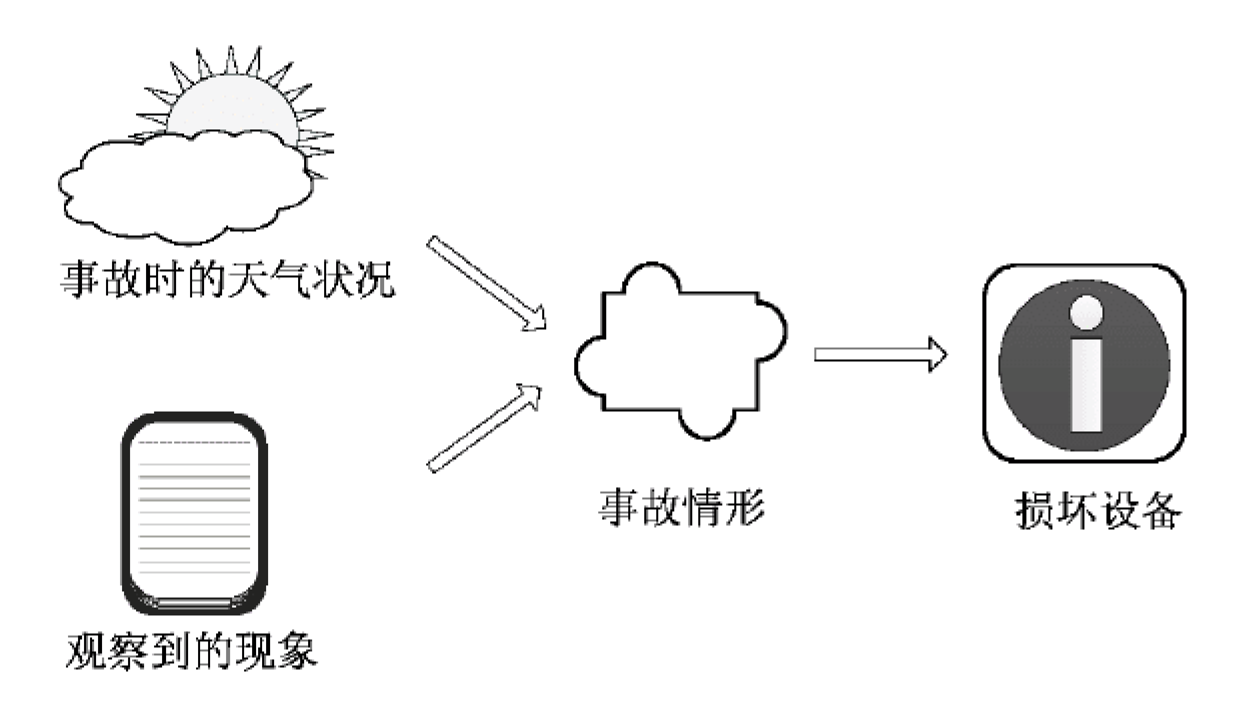


图 7.14　推理逻辑

有了初始网络图后,请专家增删证据与假设间的连接以修正网络图,最后再将修正后的网络图与专家的心智模式加以比对,以确认架构出的贝叶斯推论网络符合专家真正的推理过程;也就是节点间的每一个箭头连结关系,都符合专家进行事故诊断时使用的因果推理。经过与专家的讨论修正后,即可确定贝叶斯推论网络的架构,进一步给定推理所需要的各式参数。

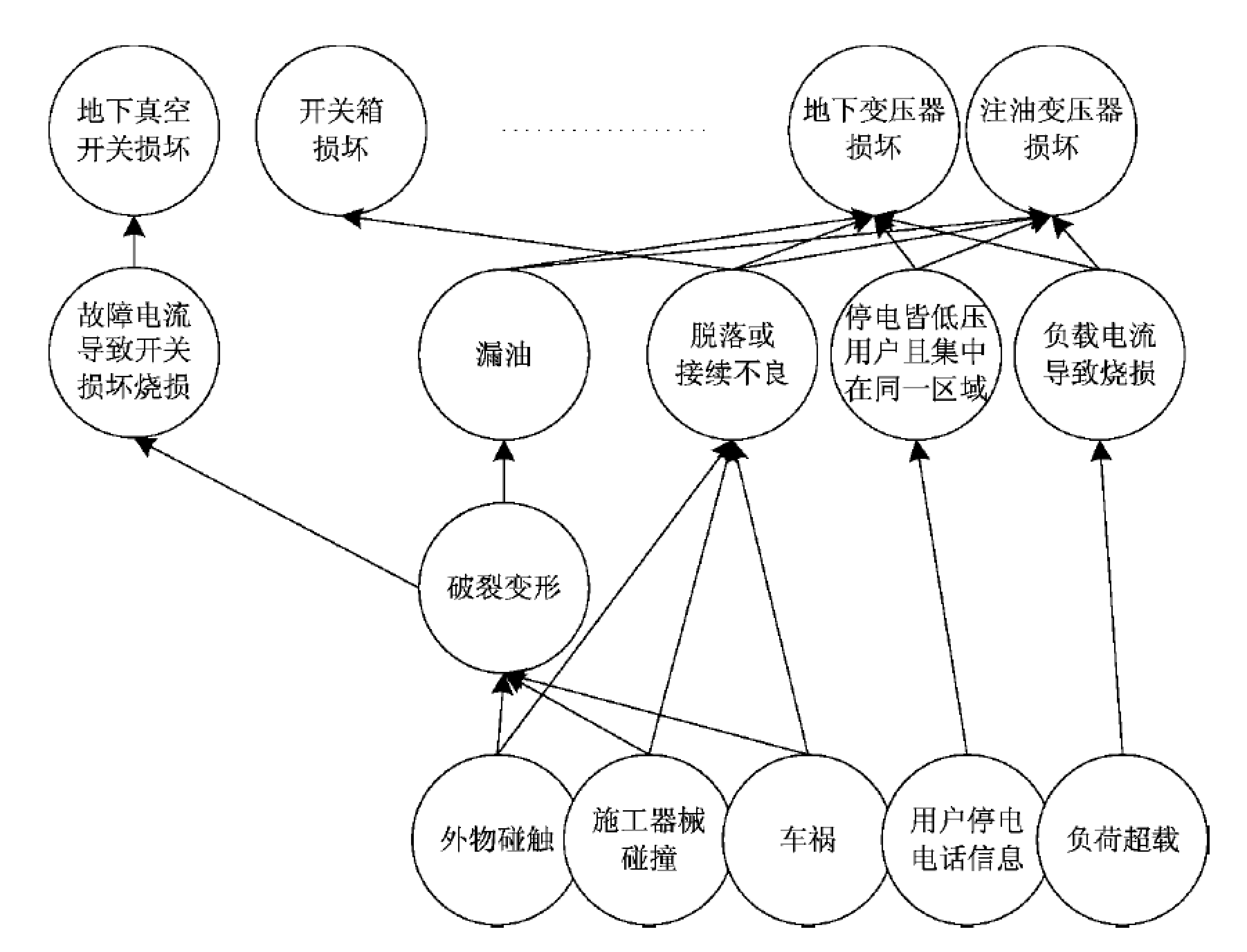


图 7.15　地下线路系统事故定位贝叶斯网络图的局部

7.5.3　给定贝叶斯推理网络的参数

在贝叶斯推理网络架构确定后，为了提取专家知识以计算似然函数，必须同时借助数据分析与口语辨证的方法。贝叶斯网络中每一个推论规则(即 if E_i then H)都包含四个参数：$P(H)$、$P(E_i)$、λ_i 与 $\bar{\lambda}_i$，研究小组借由历史数据估计各个节点的先验概率 $P(H)$，以及与目标假设相关的各种证据之先验概率 $P(E_i)$。另一方面，研究小组为了求得 λ_i 与 $\bar{\lambda}_i$，数次与专家进行结构性访谈并取得相关数据。对于每一个推理法则，必须询问专家四个相对的条件概率，即 $P(E_i|H)$、$P(E_i|\overline{H})$、$P(\overline{E_i}|H)$、$P(\overline{E_i}|\overline{H})$。举例而言，表 7.10 的访谈得到四个相对概率，例如，当注油变压器发生故障时，会观察到漏油事件的概率为 20%，即可根据这些判断求得观察到漏油的胜算比 λ_i 与没有观察到漏油的胜算比 $\bar{\lambda}_i$，计算如下：

$$\lambda_i=\frac{P(E_i|H)}{P(E_i|H)}=\frac{0.20}{0.10}=2$$

$$\bar{\lambda}_i=\frac{P(\bar{E}_i|H)}{P(E_i|H)}=\frac{0.80}{0.90}=0.89$$

表 7.10　比较下列四种状况并给予发生的相对概率值

状　况		相对概率
(1) 当注油变压器发生故障，会观察到漏油事件的概率	$P(E_i\|H)$	20%
(2) 当其他设备发生故障(不包含注油变压器)，会观察到漏油事件的概率	$P(E_i\|\overline{H})$	10%
(3) 当注油变压器发生故障，不会观察到漏油事件的概率	$P(\overline{E_i}\|H)$	80%
(4) 当其他设备发生故障，不会观察到漏油的概率	$P(\overline{E_i}\|\overline{H})$	90%

依据上述的访谈与计算，逐一得出整个贝叶斯推理网络需要的所有参数，进而验证条件独立的假设是否成立以确保贝叶斯推理网络的效度。因此，研究小组请专家评估，当一条推理法则改变时，其他指向同一个假设的推理法则是否也会改变。若专家认为其他推论法则也会改变，就必须修正网络图，直到所有的推论法则都满足条件独立的假设为止。发展推论法则过程如同构建贝叶斯网络图一般，都是不断反复的过程，经过专家知识的提取、译码、推导以及调整的步骤后，即完成最终的贝叶斯推理网络。

7.5.4 验证贝叶斯推理网络

贝叶斯推理网络的信度与效度是本系统重要的衡量指针。贝叶斯推理网络架构系经由专家的反复确认与修正，所以可以提高信度。而要验证贝叶斯推理网络的效度，本案例采用历史数据作为效标，将预测的结果与历史数据作比对，并计算两者之间的相关系数，以推论其效标效度(criteria-related validity)。亦即经由专家提供而计算得到的 λ_i 与 $\bar{\lambda}_i$，配合贝叶斯推论所得到的验后概率去预测最有可能损坏的设备。

为验证地下线路系统事故定位的贝叶斯网络，本案例以最常造成事故的三个原因："雨天"、"自然劣化"和"施工器械碰触"，作为检验贝叶斯推理网络效度的证据。选用了三年间该电力公司在某一区域的停电事故记录，样本数共 767 笔数据，经统计可得某原因发生时(例如雨天)，各种设备发生故障的实际概率值。再将该原因输入贝叶斯网络后可得各种设备发生故障的推论概率值。推论概率值与实际概率值的相关系数如表 7.11 所列，相关系数显示两者呈现高度正相关。换言之，本案例构建的贝叶斯推理网络具有良好的效度。限于篇幅，仅列出事故原因为"雨天"情况下的验证结果于图 7.16。

表 7.11 推论与实际值的相关系数(Pearson 相关系数，使用双尾检定)

情　况	样　本　数	相关系数
自然劣化	447	0.985
施工器械碰触	111	0.980
雨天	209	0.830

7.5.5 案例小结

贝叶斯网络的构建为借由提取大量的专家知识，以模仿实际配电馈线中各种造成事故的因素之间的因果关系，因此构建完成的贝叶斯推理网络除了协助事故定位外，亦可以作为事故定位专家系统的知识库与法则库。

本案例所发展的解决方案可以模拟多种配电馈线事故的情境作为新进人员的训练教材、分别建立架空与地下线路的事故定位统计推理模型架构，使各区处间得以交换经验。此外，统计数据分析或专家访谈亦有助于设计或修改现行的事故停电记录表，以及通过事故停电统计数据的整理，发掘特定馈线敷设(容易导致馈线事故)的盲点。

实际数据验证的结果显示在配电高压馈线事故定位中，贝叶斯网络非常具有潜在应用价值。当发生停电事故时，电力公司人员可借由构建出的贝叶斯网络，输入其观察到的现象，例如天气状况、施工状况等，快速地将可能发生故障的设备锁定在有限范围内，并借由推

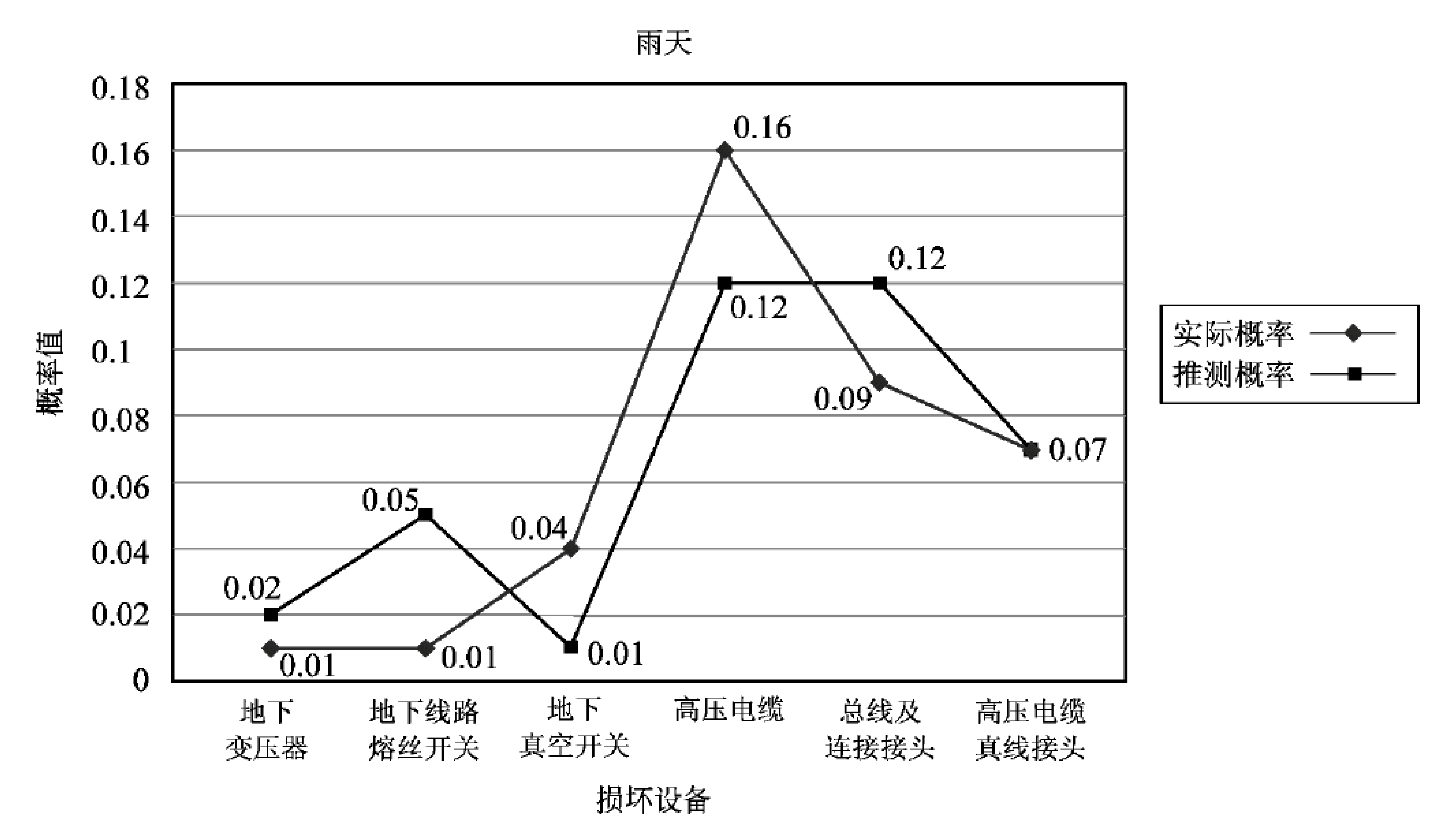


图 7.16　在雨天状况下推论值与实际数据比较

论得到的概率值将可能发生故障的设备加以排序，即由最可能发生故障的设备开始查起，以降低故障排除的时间。

7.6　结论

贝叶斯网络可以通过分析历史数据、结合主观概率与贝叶斯推论，以建立结合统计决策理论、实证数据和专家判断的数据挖掘方法。由于贝叶斯推论在抽样信息不足时，亦可以利用先验概率来计算未来风险，因此不会因为数据不足而面临无法分析的困难。贝叶斯推论亦可通过似然函数的计算来修正先验概率，并以所得的验后概率来进行风险评估与决策制定。大部分贝叶斯推论的相关研究都会选择与似然函数共轭的先验概率，以便推导验后概率分布。因此，贝叶斯推论架构也可以说是针对数据之本质去选择合适的先验概率与似然函数，使数据特性与贝叶斯推论模式相符合，并通过对先验概率分布的修正与验证来获得有效的决策模式。

随着科技的进步以及数据挖掘技术的发展，管理者得以自动或半自动方式从大量数据中搜索出有用的信息，并配合贝叶斯分类架构进行推论，从复杂且高维度的数据中找出显著的分类规则，并建立新观察点的分类模式，以降低风险并提高决策的正确性。

问题与讨论

1. 某房产销售搜集 1000 位 35 岁的工程师的婚姻状态与购置房屋不动产的统计数据，其中 500 位工程师已婚，且共有 200 位已购买不动产，另外 500 位单身的工程师中，有 400 位尚未购买不动产，所有检验统计数据列于下页表。

购买房屋不动产与婚姻状态的统计表

婚姻状态 购买不动产	单　身	已　婚	合　计
已购买(H_1)	100	200	300
未购买(H_2)	400	300	700
合计	500	500	1000

(1) 请计算来自已婚的工程师中已购买房屋不动产的条件概率。

(2) 请计算来自单身的工程师中尚未购买房屋不动产的条件概率。

(3) 若今天有一位已婚的工程师，请预测该位工程师是否已经购买房屋不动产？

2. 下表为天气状况与张三、李四与王五带伞出门状况的统计表。请根据统计数据回答以下问题。

编 号	天　气	张三带伞	李四带伞	王五带伞
1	晴	否	否	是
2	晴	否	否	否
3	阴	是	否	是
4	阴	是	是	否
5	雨	是	是	否
6	雨	是	是	是
7	阴	是	否	否
8	雨	是	是	是
9	晴	否	否	否
10	晴	否	否	是

(1) 假设在未来几天以后，在未观察当天天气的情况，且不知道张三、李四、王五带伞情况之下，分析者该如何以贝叶斯分析的角度描述当天天气的状况？

(2) 承上题，若分析者已知张三带伞出门，则其对天气描述状况应修正为何？

(3) 承题(1)，若分析者已知李四带伞出门，则其对天气描述状况应修正为何？

(4) 承题(1)，若分析者已知王五带伞出门，则其对天气描述状况应修正为何？

3. 承上题，假设已知张三带伞，李四没带伞，试以朴素贝叶斯分类法推测天气的分布状况，并讨论在此情况下朴素贝叶斯分类法的适用性，并述明原因为何。再者，除了朴素贝叶斯分类法以外，是否能由其他角度来推测当天的天气状况？若有，请比较其与朴素贝叶斯分类法的优缺点。

4. 在题 2 的数据中，请比较张三、李四与王五的带伞状况对预测天气状况的贡献度。

5. 在过去的经验中，发现越来越多高龄人口有驼背的困扰，假设与驼背相关的属性有“年龄”、“身高”、“性别”，若某医学中心搜集了 10 笔病患的个人数据，如下表：

编号	年龄/岁	身高/cm	性别	驼背 D	编号	年龄/岁	身高/cm	性别	驼背 D
1	>50	>175	男	是	6	>50	≤175	女	是
2	≤50	>175	男	否	7	≤50	≤175	男	否
3	>50	≤175	女	否	8	≤50	≤175	女	否
4	≤50	≤175	女	否	9	≤50	>175	男	是
5	>50	≤175	男	否	10	≤50	≤175	女	否

(1) 根据以上数据，如有一位病患的个人数据为“年龄>50、身高>175、男性”，请利用朴素贝叶斯分类法预估该位病患是否会有驼背？

(2) 承上题，如有另一位病患的个人数据为“年龄≤50、身高≤175、女性”，则其是否会有驼背？

6. 某医院有三种检测受试者是否罹患 AIDS 的检测方法，并已知三种检测方法在受测者有罹患 AIDS 与无罹患 AIDS 之下检查结果显隐性的概率分布如下表所示。假设在患者是否罹病的情况确定之下，三种检测的结果可视为独立，试回答下列问题。

(1) 假设未进行检测前，医生依据经验对某受试者甲罹患 AIDS 的概率推断(先验信息)为 0.5。之后进行检测 1 结果为显性，请问此时医生对甲罹患 AIDS 的概率应修正为何？

(2) 承上题，假设之后继续进行检测 2，结果仍为显性，请问此时医生对甲罹患 AIDS 的概率应修正为何？

(3) 承上题，假设之后继续进行检测 3，结果为隐性，请问此时医生对甲罹患 AIDS 的概率应修正为何？

(4) 假设未进行检测前，医生依据经验对某受试者乙罹患 AIDS 的概率推断(先验信息)为 0.01。之后进行三个检测的结果皆呈现显性，请问此时医生对乙罹患 AIDS 的概率应修正为何？

是否患病	检查结果	检测 1	检测 2	检测 3
是	显性	0.90	0.99	0.95
	隐性	0.10	0.01	0.05
否	显性	0.10	0.20	0.25
	隐性	0.90	0.80	0.75

7. 令 Y 为一随机变量，其概率密度函数如下：

$$P(Y=y)=\begin{cases}0.2, & y=\theta \\ 0.8, & y=\theta+1\end{cases}$$

其中，θ 为参数，且先验分布为

$$P(\theta=t \mid \beta=1)=\begin{cases}0.9, & t=1 \\ 0.1, & t=2\end{cases}$$

$$P(\theta = t \mid \beta = 2) = \begin{cases} 0.5, & t = 2 \\ 0.5, & t = 3 \end{cases}$$

β 的先验分布为

$$P(\beta = t) = \begin{cases} 0.5, & t = 1 \\ 0.5, & t = 2 \end{cases}$$

请回答下列问题。

(1) 请问 θ 与 β 的先验期望值 $E[\theta]$ 与 $E[\beta]$ 为何?

(2) 假设观察到 $Y=1$,θ 与 β 的验后期望值 $E[\theta|Y=1]$ 与 $E[\beta|Y=1]$ 为何?

(3) 假设观察到 $Y=2$,θ 与 β 的验后期望值 $E[\theta|Y=2]$ 与 $E[\beta|Y=2]$ 为何?

(4) 假设观察到 $Y=4$,θ 与 β 的验后期望值 $E[\theta|Y=4]$ 与 $E[\beta|Y=4]$ 为何?

8. 某工厂工程师从生产历史数据中搜集了 15 笔产品的制造加工机台(M)与最终的检测结果(Y),如果检测结果为不良品则标示为 1,检测结果为良品则为 0,以第一笔数据为表示该产品加工的机台顺序为 $M_2 \to M_3 \to M_6$,如下表:

训练数据集

编号	M_1	M_2	M_3	M_4	M_5	M_6	M_7	Y
1	0	1	1	0	0	1	0	0
2	0	1	0	0	1	0	1	0
3	0	1	0	0	1	0	1	0
4	1	0	0	1	0	0	1	1
5	0	1	1	0	0	0	0	0
6	1	0	0	0	0	0	1	1
7	1	0	0	1	1	0	1	0
8	1	0	1	0	0	1	0	0
9	0	1	0	1	0	1	0	0
10	0	1	0	1	0	1	0	1
11	1	0	1	0	0	1	0	0
12	1	0	1	0	0	1	0	0
13	1	0	1	0	0	0	1	1
14	1	0	0	1	0	0	1	1
15	0	1	0	1	0	1	0	1

(1) 请根据上表画出贝叶斯网络的网络结构图。

(2) 若某产品加工的机台顺序为 $M_2 \to M_5 \to M_6$,试预测该产品可能检测结果?

(3) 若某产品加工的机台顺序为 $M_1 \to M_4 \to M_7$,试预测该产品可能检测结果?

第 8 章

粗糙集理论

8.1 粗糙集理论

粗糙集理论(rough set theory,RST)是一种处理数据分类的数据挖掘方法。当数据属于定性数据(qualitative data)或不确定性(uncertainty)数据,无法使用一般的统计方法时,粗糙集理论可以在信息不完整(incomplete)和信息不一致(inconsistent)下,用来归约数据集合、发掘隐藏的数据样型和数据相关性,以产生有用的分类规则(Tseng *et al.*, 2004; Pawlak, 1982, 1991)。例如,Chien 和 Chen(2007)应用粗糙集理论提取员工在工作表现、工作年资与辞职原因之间的关系,以协助案例公司挑选适合的人才,也有助于发展新的人才挑选策略;Chien 等(2014)应用粗糙集理论,分析用户经验的问卷调查数据,以挖掘 3C 产品设计的参考规则。

8.2 粗糙集理论基本概念

粗糙集理论运算过程的符号及定义如下:

S	表示信息系统,$S=(U,A,V,f)$
U	宇集合,对象 x_j 的有限集合,$x_j \in U$
A	表示一个包含有属性 a_k 的有限集合,$a_k \in A$
V	属性值 V_{a_k} 的宇集合,$V=\bigcup_{a_k \in A} V_{a_k}$
f	信息函数,$x_j \in U, a_k \in A, f(x_j, a_k) \in V_{a_k}$
X	宇集合 U 的部分集合,$X \subseteq U$
D	属性集合 A 的非空子集合,$a_d \in D$
V_{a_k}	属性 a_k 的属性值,$a_k \in A$
I_D	D 的不可分辨关系
$I_D(.)$	不可分辨关系 D 下的基本集
$U\|D$	表示在 D 的等价关系中的所有基本集所成的集合
$\underline{D}(X)$	下近似集合,表示所有在宇集合里属性集合 D 中的等价关系对象可以完全包含在集合 X 中
$\overline{D}(X)$	上近似集合,表示在宇集合里属性集合 D 中的等价关系对象可能被包含在集合 X 中
$BN_D(X)$	边界区,表示在现有信息下(属性集合 R)无法明确地分类到属于集合 X 中,或不属于集合 X 中,$BN_D(X)=\overline{D}(X)-\underline{D}(X)$

$\alpha_D(X)$　　近似集合 X 的准确率，$0 \leqslant \alpha_R(X) \leqslant 1$

$pos_D(X)$　　正域 $pos_R(X)$ 为根据属性集合 R 下，宇集合 U 中能完全确定归属于集合 X 元素的集合，$pos_R(X)=\underline{R}X$

$Reduct(D)$　　包含 D 所有 reduct 的集合

$Core(D)$　　集合 D 的核，在 D 上，U 中所有不可省略关系的集合

8.2.1 信息系统与决策表

信息系统(**information systems**)包含四种组成元素：一个属于有限集合的宇集合(universe)，一个属于有限集合的属性(attributes/features)集合，对象在每一个特性里所表现的值(value)，以及代表所有对应关系的决策函数或称信息函数(function)。可以用符号表示如下：

$$S=(U,A,V,f) \tag{8.1}$$

其中，S 表示此信息系统；U 表示对象(objects)x_j 的非空有限集合，$x_j \in U$；A 表示属性 a_k 的非空有限集合，$a_k \in A$；$V=\bigcup_{a_k \in A} V_{a_k}$，$V_{a_k}$ 是属性 a_k 所代表的值；函数表示为 $f:U\times A \to V$，指定宇集合 U 中每个对象 u_j 的属性，对所有的 $x\in U$，$a_k \in A$ 使得 $f(u_j,a_k)\in V_{a_k}$。

决策表(**decision table**)是呈现当数据满足哪些条件下，会产生的决策之间的因果关系。例如，表 8.1 配电事故诊断记录数据的决策表来说明粗糙集理论的信息系统，其中**条件属性**(**condition features**)有 3 个，分别是天气、事故情形、事故原因；**决策属性**(**decision feature**)为损坏部位。换言之，宇集合为这 5 笔事故诊断数据 $U=\{x_1,x_2,x_3,x_4,x_5\}$共 5 个物件；属性集合包含三个条件属性以及一个决策特性，$A=\{C\cup D\}$，$C=\{$天气，事故情形，事故原因$\}$，$D=\{$损坏部位$\}$；属性值分别表示于各属性之下，例如 $V_{天气}=\{0$：雨天，1：阴天$\}$；函数的对应关系例如对象 1(x_1)其天气为雨天、事故情形为烧断、事故原因为自然劣化、损坏部位为高压电缆，所以对象 1 与天气所对应的值为 0、事故情形对应的值为 1、事故原因对应的值为 1、损坏部位对应的值为 0。

表 8.1　事故诊断记录数据的决策表

编　号	条件属性			决策属性
	天气	事故情形	事故原因	损坏部位
	0：雨天 1：阴天	0：挖断 1：烧断 2：断落	0：外物碰触 1：自然劣化	0：高压电缆 1：熔丝链开关
x_1	0	1	1	0
x_2	1	0	0	0
x_3	1	2	1	1
x_4	0	1	1	1
x_5	1	1	0	1

8.2.2 等价关系

等价关系(equivalence relations)是当分析一组数据时，若对象与对象之间因为在某些

属性上包含相同信息，而变成难以辨别(indiscernibility)的关系，则称此两个对象有等价关系且属于同一个分类的交集(Pawlak，1991)。若属性集合 D 为属性集合 A 的非空子集合，$D\subseteq A$，则可定义对象 x_1 与 x_2 的不可分辨关系如下：

$$(x_1,x_2)\in I_D \Leftrightarrow f(x_1,a_d)=f(x_2,a_d),\forall a_d\in D \tag{8.2}$$

以表 8.1 数据为例，可以将部分条件属性的等价关系表示如下：

I_D(天气) $=\{0:\{x_1,x_4\},1:\{x_2,x_3,x_5\}\}$

I_D(事故情形) $=\{0:\{x_2\},1:\{x_1,x_4,x_5\},2:\{x_3\}\}$

I_D(事故原因) $=\{0:\{x_2,x_5\},1:\{x_1,x_3,x_4\}\}$

I_D(天气，事故原因) $=\{\{0,1\}:\{x_1,x_4\},\{1,0\}:\{x_2,x_5\},\{1,1\}:\{x_3\}\}$

I_D(天气，事故情形，事故原因) $=$

$\{\{0,1,1\}:\{x_1,x_4\},\{1,0,0\}:\{x_2\},\{1,2,1\}:\{x_3\},\{1,1,0\}:\{x_5\}\}$

8.2.3　近似空间

近似空间(approximation space)是由 N 个对象的宇集合与属性集合的等价关系构成。在一个属性集合的等价关系中，等价类(equivalence class)形成基本集(elementary sets)。$U|D$ 表示在 D 的等价关系中的所有基本集所成的集合。例如，事故情形这个属性的等价关系中，有 $\{x_2\}$、$\{x_1,x_4,x_5\}$、$\{x_3\}$ 等三个基本集，表示为 $U|$事故情形$=\{\{x_1,x_4,x_5\},\{x_2\},\{x_3\}\}$。

粗糙集并以“**下近似**”(**lower approximation**)和“**上近似**”(**upper approximation**)两个集合来表现数据的不确定性。假设集合 X 是宇集合 U 的部分集合，D 为某一属性集合，则定义下近似 $\underline{D}(X)$ 与上近似 $\overline{D}(X)$ 如下：

$$\begin{aligned}\underline{D}(X)&=\{x\in U:I_D(x)\subseteq X\}\\&=\bigcup\{Y\in U\mid D:Y\subseteq X\}\end{aligned} \tag{8.3}$$

$$\begin{aligned}\overline{D}(X)&=\{x\in U:I_D(x)\cap X\neq\varnothing\}\\&=\bigcup\{Y\in U\mid D:Y\cap X\neq\varnothing\}\end{aligned} \tag{8.4}$$

其中，Y 为基于属性集合 D 下的描述，I_D 表示属性集合 D 的基本集。因此，集合 X 的下近似表示所有在宇集合里属性集合 D 中的等价关系对象可以完全被包含在集合 X 中；而上近似则表示在宇集合里属性集合 R 中的等价关系对象可能被包含在集合 X 中。另外，定义**边界区**(**boundary region**)为

$$BN_D(X)=\overline{D}(X)-\underline{D}(X) \tag{8.5}$$

边界区 $BN_D(X)$ 表示在边界区里的对象，在现有信息下无法明确的将它分类到属于或不属于集合 X 中。正域 $pos_D(X)$ 为根据属性集合 D 下，宇集合 U 中能完全确定归属于集合 X 元素的集合，如式(8.6)；负域 $neg_D(X)$ 表示根据属性集合 D 下，宇集合 U 中确定无法归属于集合 X 元素的集合，如式(8.7)。正域与负域的关系如图 8.1 所示。

$$pos_D(X)=\underline{D}(X) \tag{8.6}$$

$$neg_D(X)=U-\overline{D}(X) \tag{8.7}$$

以表 8.1 为例，若属性集合 D 为天气与事故情形，属性集合 D 的等价关系中的所有元素所成的集合为 $U|$\{天气，事故情形\}$=\{\{x_1,x_4\},\{x_2\},\{x_3\},\{x_5\}\}$，假设集合 X 代表损坏部位为熔丝链开关，所以集合 X 为 $X=\{x_3,x_4,x_5\}$，因此，集合 X 的下近似为 $\underline{D}X=\{x_3,$

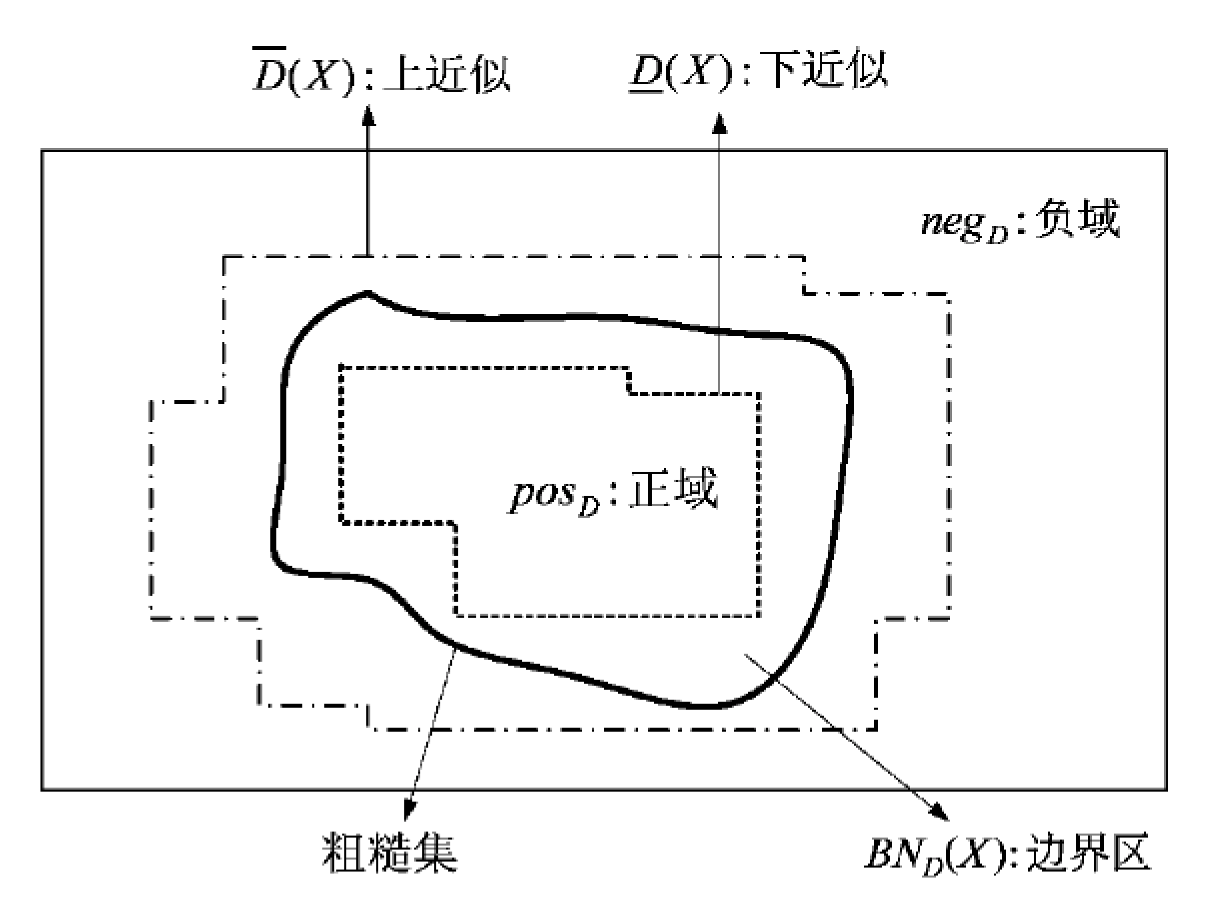


图 8.1　粗糙集示意图

$x_5\}$,而集合 X 的上近似为 $\overline{D}X=\{x_1,x_3,x_4,x_5\}$,边界区将上近似集合减去下近似集合,得到 $BN_D(X)=\overline{D}(X)-\underline{D}(X)=\{x_1,x_4\}$,表示在运用天气与事故情形这两个属性信息下,可以知道对象 3 与对象 5 属于集合 X,而在现有天气与事故情形这两个信息下,无法明确判断对象 1 与对象 4 是属于集合 X 还是不属于集合 X,如图 8.2 所示,而正域为 $pos_D(X)=\{x_3,x_5\}$,负域 $neg_D(X)=\{x_2\}$。

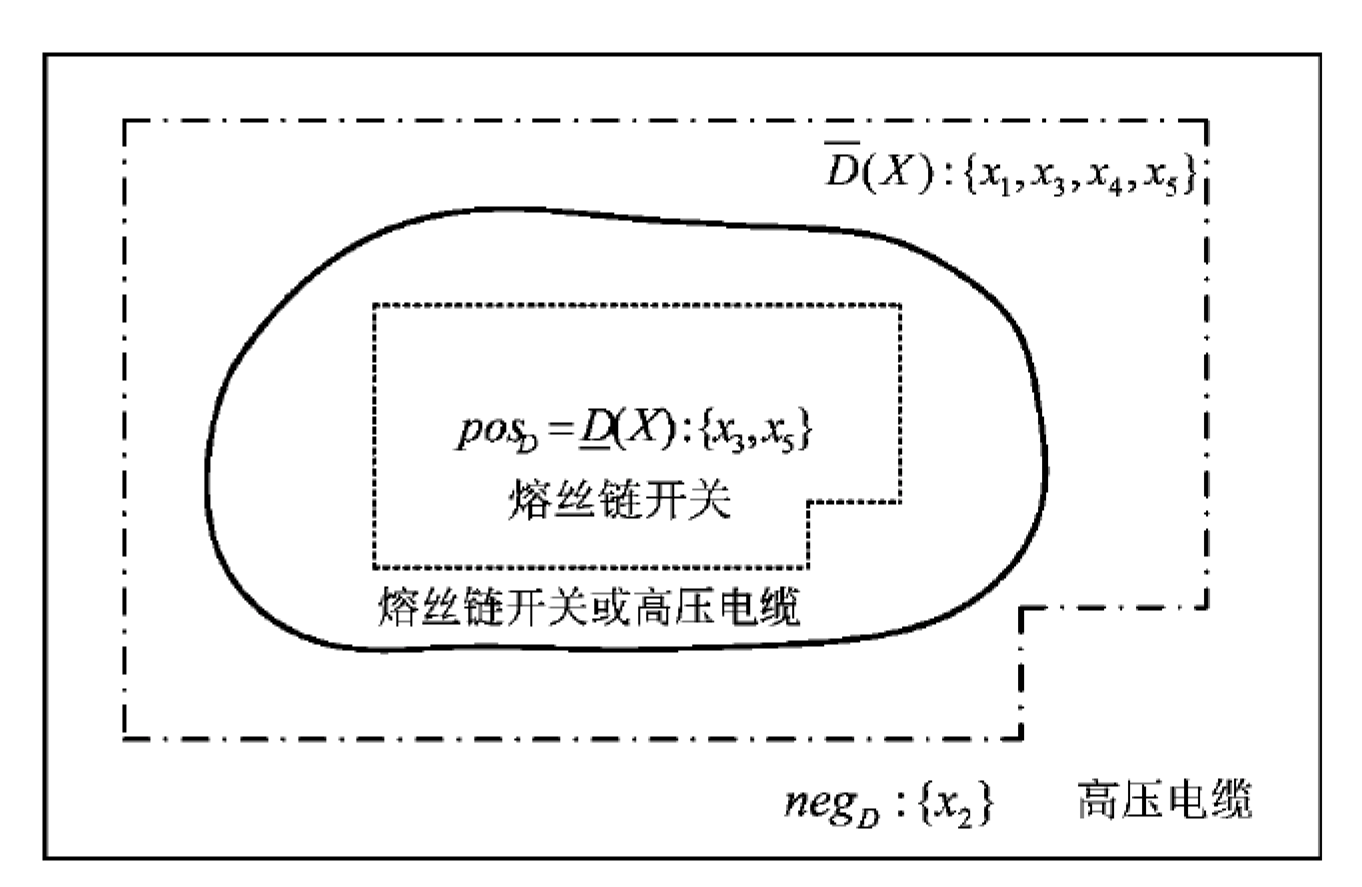


图 8.2　下近似与上近似说明图

8.2.4　近似集合的准确率

一个集合若在 D 的等价关系信息下存在边界区,则表示此集合在 D 的信息下仍有无法明确分类属于集合中或不属于集合中的对象存在,因此边界区的元素越多,则表示此近似集合的准确程度越低。近似集合 X 的**准确率(accuracy)**可定义为

$$\alpha_D(X)=\frac{card\,\underline{D}(X)}{card\,\overline{D}(X)} \tag{8.8}$$

其中,$X\neq\varnothing$,$card(\cdot)$表示集合中对象的个数。准确率 $\alpha_D(X)$表示对于集合 X,利用属性集合 D 的等价关系,对象是否能够准确分类到集合的程度。当 $\alpha_D(X)=1$ 时,即边界区为空集合,则集合 X 称为属性集合 D 可限定的(D-definable)集合,或称精确(precise/crisp)集合;假如

$\alpha_D(X)<1$，即边界区不是空集合，则集合 X 称为属性集合 D 不可限定的(D-undefinable)集合，或称为粗糙(rough)集合。

以表 8.1 为例，若属性集合 D 为天气与事故情形，则 $U|D=\{\{x_1,x_4\},\{x_2\},\{x_3\},\{x_5\}\}$，假设集合 X 的损坏部位为熔丝链开关，即 $X=\{x_3,x_4,x_5\}$，得到集合 X 的下近似 $\underline{D}X=\{x_3,x_5\}$，上近似 $\overline{D}X=\{x_1,x_3,x_4,x_5\}$。因此，准确率 $\alpha_D(X)=2/4$。

8.2.5　分类的准确率与属性相依程度

粗糙集理论可以作为一种处理数据分类的决策方法，衡量该属性集合 D 能否解释对应分类目标，可借由上近似集合与下近似集合的比率来定义。假设 $F=\{X_1,X_2,X_3,\cdots,X_n\}$，$X_i=\{x_1,x_2,x_3,\cdots,x_k\}$，是一组含有 n 个非空集合所组成的集合，每一个集合 X 代表一个分类类型，此组集合的下近似为 $\underline{D}(F)=\{\underline{D}(X_1),\underline{D}(X_2),\underline{D}(X_3),\cdots,\underline{D}(X_n)\}$，上近似为 $\overline{D}(F)=\{\overline{D}(X_1),\overline{D}(X_2),\overline{D}(X_3),\cdots,\overline{D}(X_n)\}$，此组集合 X 可被属性集合 D 定义为

$$\alpha_D(F)=\frac{\sum card\ \underline{D}(X_i)}{\sum card\overline{D}(X_i)} \tag{8.9}$$

另外，可定义此组近似集合 D 对分类属性集合的相依程度为

$$\gamma_D(F)=\frac{\sum card\ \underline{D}(X_i)}{cardU}=\frac{|pos_D(F)|}{|U|} \tag{8.10}$$

$\alpha_D(F)$即表示在利用属性集合 D 的等价关系信息下分类对象，有多少比率可完全由属性集合 D 解释；而 $\gamma_D(F)$相依程度即表示在属性集合 D 的等价关系信息下，宇集合对象能被正确划分到集合 F 中的比率，$pos_D(F)$表示在属性集合 D 下，所有属性集合 F 的正域，即为所有分类集合属性下的下限集合 $\underline{D}(X)$ 的联集。

以表 8.1 为例，属性集合 D 为天气与事故情形，则 $U \mid D=\{\{x_1,x_4\},\{x_2\},\{x_3\},\{x_5\}\}$，假设集合 $F=\{X_1:\{x_1,x_2\},X_2:\{x_3,x_4,x_5\}\}$，即表示集合 X_1 为损坏部位为高压电缆，集合 X_2 为损坏部位为熔丝链开关。集合 X_1 的下近似为 $\underline{D}(X_1)=\{x_2\}$，集合 X_1 的上近似为 $\overline{D}(X_1)=\{x_1,x_2,x_4\}$，集合 X_2 的下近似为 $\underline{D}(X_2)=\{x_3,x_5\}$，集合 X_2 的上近似为 $\overline{D}(X_2)=\{x_1,x_3,x_4,x_5\}$。因此，分类的准确率 $\alpha_D(F)=(1+2)/(3+4)=3/7$；属性 D 与集合 F 的相依程度 $\gamma_D(F)=3/5$。换言之，在利用天气与事故情形的等价关系信息下分类对象，可以正确分类属于高压电缆损坏或熔丝链开关损坏的比例为 3/7；而属性相依程度即表示在利用天气与事故情形的等价关系信息下分类对象，这五个对象能正确分类属于高压电缆损坏或熔丝链开关损坏的比例为 3/5。

8.2.6　简化

粗糙集理论以**简化(reducts)**来表示属性归约后的集合，以代表在条件属性集合中的最小充分子集合。也就是说，利用整个条件集合信息所分类的结果，与利用简化集合信息所分类的结果相同。假如属性集合 D 是属性集合 A 的子集合，$D\subseteq A$，且 $a_d\in D$，则

$$I_D=I_{D-\{a_d\}} \tag{8.11}$$

表示属性 a_d 在属性集合 D 中是相依的(dispensable)属性；否则，属性 a_d 在属性集合 D 中则为独立的(indispensable)属性。假如一个集合 E 是独立的属性集合，$E\subseteq D$，且 $I_E=I_D$，则

称 E 是 D 的 reduct。所以一个属性集合可能包含多个 reduct。

以表 8.1 来说明 reduct,其中属性集合 D 是天气、事故情形与事故原因,下列为不同等价关系下的基本集:

$$U=\{x_1,x_2,x_3,x_4,x_5\}$$

$$U\mid D=U\mid \text{天气,事故情形,事故原因}=\{\{x_1,x_4\},\{x_2\},\{x_3\},\{x_5\}\}$$

$$U\mid(D-\text{天气})=\{\{x_1,x_4\},\{x_2\},\{x_3\},\{x_5\}\}=U\mid D$$

$$U\mid(D-\text{事故情形})=\{\{x_1,x_4\},\{x_2,x_5\},\{x_3\}\}\neq U\mid D$$

$$U\mid(D-\text{事故原因})=\{\{x_1,x_4\},\{x_2\},\{x_3\},\{x_5\}\}=U\mid D$$

由于 $U|(D-\text{事故情形})\neq U|D$,所以事故情形是独立的属性,而 $U|(D-\text{天气})=U|D$ 与 $U|(D-\text{事故原因})=U|D$,所以天气与事故原因是相依的属性,因此,属性集合 D 的 reduct 有{事故情形、事故原因}以及{天气、事故情形}两个。

粗糙集理论可用来简化属性以产生关键属性,并进而简化相等的类组以发掘数据中的决策规则。假设有一组对象集合 F,$F=\{X_1,X_2,\cdots,X_n\}$,$X_i\subseteq U$,且有一个宇集合的子集合 Y,$Y\subseteq U$,使得 $\cap F\subseteq Y$。若 $\cap(F-\{X_i\})\subseteq Y$,则称集合 X_i 在集合 F 的交集中是与集合 Y 相依的,否则集合 X_i 在集合 F 的交集中是与集合 Y 独立的。当集合 F 的一组子集合 H,在集合 F 的交集中是与集合 Y 独立的,且 $\cap H\subseteq Y$,则称集合 H 是集合 Y 的 reduct。因为一组集合可以产生多个 reduct,而找到最少的 reduct 是一个 NP-hard 问题,帕夫拉克(Pawlak,1991)提出 reduct 的产生程序,包含以下四个步骤:

步骤 0:将数据集合物件由 $1\sim n$ 编号,从 $i=1$ 开始。

步骤 1:若有 m 个条件属性,则在对象 i 中,产生由 $1\sim m-1$ 个条件属性所组成的 reduct。

步骤 2:令 $i=i+1$,假如所有的对象都已经计算过,则进入步骤 3;否则回到步骤 1。

步骤 3:搜集所有产生的 reduct。

利用表 8.1 的数据来产生对象集合的 reduct,由于要推导损坏部位与其他三项属性的规则,所以形成高压电缆 $=\{x_1,x_2\}$ 与熔丝链开关 $=\{x_3,x_4,x_5\}$ 两个决策集合。以下以对象 1 与对象 2 的推导过程为例说明:

步骤 1:当 $i=1$ 时,

当 $m=1$ 时,

$[0]_{\text{天气}}=\{x_1,x_4\}$;

$[1]_{\text{事故情形}}=\{x_1,x_4,x_5\}$;

$[1]_{\text{事故原因}}=\{x_1,x_3,x_4\}$。

当 $m=2$ 时,

$[0,1]_{\text{天气、事故情形}}=\{x_1,x_4\}$;

$[0,1]_{\text{天气、事故原因}}=\{x_1,x_4\}$;

$[1,1]_{\text{事故情形、事故原因}}=\{x_1,x_4\}$。

由于对象 1 的简化集合都未包含于对象 1 的决策集合高压电缆 $=\{x_1,x_2\}$ 中,所以对象 1 未产生 reduct。

步骤 2:当 $i=1+1$ 时,回到步骤 1;

步骤 1:当 $i=2$ 时,

当 $m=1$ 时，

$[1]_{天气}=\{x_2,x_3,x_5\}$；

$[0]_{事故情形}=\{x_2\}$；

$[0]_{事故原因}=\{x_2,x_5\}$。

当 $m=2$ 时，

$[1,0]_{天气、事故情形}=\{x_2\}$；

$[1,0]_{天气、事故原因}=\{x_2,x_5\}$；

$[0,0]_{事故情形、事故原因}=\{x_2\}$。

对象 2 的简化集合有 $[0]_{事故情形}=\{x_2\}$、$[1,0]_{天气、事故情形}=\{x_2\}$、$[0,0]_{事故情形、事故原因}=\{x_2\}$ 包含于对象 2 的决策集合高压电缆 $=\{x_1,x_2\}$ 中，所以对象 2 产生 3 个 reduct。

步骤 2：当 $i=2+1$ 时，回到步骤 1；

步骤 1：当 $i=3$ 时，

当 $m=1$ 时，

$[1]_{天气}=\{x_2,x_3,x_5\}$；

$[2]_{事故情形}=\{x_3\}$；

$[1]_{事故原因}=\{x_1,x_3,x_4\}$。

当 $m=2$ 时，

$[1,2]_{天气、事故情形}=\{x_3\}$；

$[1,1]_{天气、事故原因}=\{x_3\}$；

$[2,1]_{事故情形、事故原因}=\{x_3\}$。

对象 3 的简化集合有 $[2]_{事故情形}=\{x_3\}$、$[1,2]_{天气、事故情形}=\{x_3\}$、$[1,1]_{事故情形、事故原因}=\{x_3\}$、$[2,1]_{事故情形、事故原因}=\{x_3\}$，包含于对象 3 的决策集合熔丝链开关 $=\{x_3,x_4,x_5\}$ 中，所以对象 3 产生 4 个 reduct。

步骤 2：当 $i=3+1$ 时，回到步骤 1；

步骤 1：当 $i=4$ 时，

当 $m=1$ 时，

$[0]_{天气}=\{x_1,x_4\}$；

$[1]_{事故情形}=\{x_1,x_4,x_5\}$；

$[1]_{事故原因}=\{x_1,x_3,x_4\}$。

当 $m=2$ 时，

$[0,1]_{天气、事故情形}=\{x_1,x_4\}$；

$[0,1]_{天气、事故原因}=\{x_1,x_4\}$；

$[1,1]_{事故情形、事故原因}=\{x_1,x_4\}$。

由于对象 4 的简化集合都未包含于对象 4 的决策集合熔丝链开关 $=\{x_3,x_4,x_5\}$ 中，所以对象 1 未产生 reduct。

步骤 2：当 $i=4+1$ 时，回到步骤 1；

步骤 1：当 $i=5$ 时，

当 $m=1$ 时，

$[1]_{天气}=\{x_2,x_3,x_5\}$；

$[1]_{\text{事故情形}}=\{x_1,x_4,x_5\}$;

$[0]_{\text{事故原因}}=\{x_2,x_5\}$。

当 $m=2$ 时，

$[1,1]_{\text{天气、事故情形}}=\{x_5\}$;

$[1,0]_{\text{天气、事故原因}}=\{x_2,x_5\}$;

$[1,0]_{\text{事故情形、事故原因}}=\{x_5\}$。

对象 5 的简化集合有$[1,1]_{\text{天气、事故情形}}=\{x_5\}$、$[1,0]_{\text{事故情形、事故原因}}=\{x_5\}$包含于对象 5 的决策集合熔丝链开关$=\{x_3,x_4,x_5\}$中，所以对象 5 产生 2 个 reduct。考虑所有的对象之后，产生的 reduct 如表 8.2 所示。

表 8.2　以表 8.1 为例所产生的 reduct

Reduct 编号	天气	事故情形	事故原因	损坏部位	天气	事故情形	事故原因	损坏部位
x_1	0	1	1	0	×	×	×	×
x_2	1	0	0	0	×	0	×	0
					1	0	×	0
					×	0	0	0
x_3	1	2	1	1	×	2	×	1
					1	2	×	1
					1	×	1	1
					×	2	1	1
x_4	0	1	1	1	×	×	×	×
x_5	1	1	0	1	1	1	×	1
					×	1	0	1

从所产生的 reduct 中，可以选取有意义的决策规则，表示为：“当条件属性等于 V_{a_k} 成立，则可以推论结果为 V_d”。假设从表 8.2 所产生的 reduct 中，选取“× 0 × 0”为决策规则，其中“×”表示该属性没有包含在 reduct 中，亦即表示当事故情形等于 0 时，可以推论损坏部位为 0，所以可以得到一决策规则为“当事故情形为挖断时，可以推论损坏部位为高压电缆”(Peng，*et al*.，2004)，可与第 7 章应用贝叶斯网络所做的馈线事故定位做比较(Chien，*et al*.，2002)。

8.3　粗糙集理论产生分类规则

可以从训练数据集中，应用粗糙集理论与支持度(support)门槛产生候选规则，并利用测试数据集计算候选规则的置信度(confidence)与增益(lift)，以验证提取之候选规则作为最终分类规则。

建立候选规则之前，以随机的方式将决策表分成两组：$a\%$的数据视为训练数据组；1－



$a\%$的数据则视为测验数据组。产生分类规则的步骤如下：

(1) 定义候选规则所需之支持度门槛值 θ_s。

(2) 建立决策表与数据集。

(3) 若遇属性为连续型属性，则需经过离散化，将连续型数据分为数个区间，详细离散化方法可见第二章；否则直接进入步骤(4)。

(4) 取得训练组数据集的简化(reducts)。

(5) 根据领域专业知识判定于步骤(4)所产生的 reducts 是否合适。

(6) 根据筛选后所剩下的 reducts 组而找出规则。

(7) 输入所有训练数据集，并计算所有产生规则的支持度。若该规则支持度大于门槛值 θ_s，则应将所该规则放入候选规则集合中；若该规则的支持度小于门槛值 θ_s，则移除该规则。

(8) 直到所有规则均完成支持度门槛值的检验后，即可停止产生规则，并与领域专家讨论，剔除不符合实务的候选规则。

接着使用测验数据组验证从训练数据组所取得的候选规则，并以置信度与增益作为评选候选规则的门槛值。步骤说明如下：

(1) 设定置信度与增益门槛值，分别为 θ_c 与 θ_l。

(2) 输入所有测试数据集，以计算各候选规则的置信度与增益。

(3) 若置信度大于门槛值 θ_c，且增益大于门槛值，则此候选规则将通过测试，并作为最终分类规则；若该规则的置信度小于门槛值 θ_s，则移除该候选规则。

(4) 直到所有候选规则均完成置信度与增益的检验后，即完成产生分类规则的步骤，再与领域专家确认规则的意义。

8.4　粗糙集理论与其他分类方法的比较

表 8.3 比较四种数据挖掘方法的差异，以作为选择数据挖掘工具时的参考。在处理数据形态上，粗糙集、关联规则、决策树三种方法皆是处理分类的模式，因此可以处理的数据形态皆属于类别数据，对于数值数据的处理较为困难需要先离散化，准确率也相对较低，不易产生显著的样型。相较之下，贝叶斯网络推论主要是处理概率问题，因此可以处理离散变量或连续变量。

粗糙集理论与关联规则皆是直接从数据中挖掘出规则样型，数据并不需要假设条件，但当变量值个数太多时，应该合并成几个类别值，以增加规则的准确率；决策树方法在分支时必须根据分支方法有适当的假设条件；贝叶斯网络方法则需要假设类别条件独立。因此，粗糙集方法、关联规则、决策树皆是客观的分析数据，发现数据中有意义的样型；贝叶斯网络方法则除了客观的分析数据之外，还包含主观判断关联项目的条件概率。

在目标变量的个数限制上，这四种方法皆无法同时处理太多数量的变量，目标项可变动的变量值不能太多，否则不容易产生显著的样型规则，反而可能会产生很多杂乱的规则，必须再做进一步的筛选。贝叶斯网络方法则易因为推论的项目太多，以致无法满足类别条件独立的假设，另外，贝叶斯网络假设项目(结果事件)与证据项目(原因事件)处理数值项目时，需要事先合并与离散化，否则其推论项目太多会造成推论的困难。整体来说，这四种方

法只有贝叶斯网络可以有效地处理遗漏值，其他三种方法则需要在数据预处理阶段，先处理遗漏值的问题。如果要采取补值的方式处理遗漏值，以配电事故诊断数据的特性，应以与遗漏值同一损坏部位数据中，该属性变量值出现次数最多的值填补。

表 8.3 比较四种数据挖掘方法的差异

比较项目＼方法	粗糙集理论	关联规则	决策树	贝叶斯网络
类别数据	容易处理	容易处理	可以处理	可以处理
数值数据	必须离散化	较难处理	可以处理	可以处理
	要求高的准确率必须连续转离散	要求高的准确率必须连续转离散	要求高的准确率必须连续转离散	要求高的准确率必须连续转离散
假设条件	不需要假设条件	不需要假设条件	分支时需要假设条件	需假设类别条件独立
主观/客观	客观的处理数据	客观的处理数据	客观的处理数据，分支时需主观决定假设条件是否成立	客观的处理数据，主观的决定条件概率
方法原理	集合论（非统计方法）	含统计推论（置信度）	含统计推论（分支时检定）	统计方法之一
规则结果的解释	容易理解	容易理解	容易理解	概率表示，容易理解
目标变量个数	变量值太多无法处理（support 会太低）	变量值太多无法处理（support 会太低）	变量值太多无法处理（support 会太低）	变量值数较无限制，要求高的准确率时个数不宜太多
规则使用的属性个数	属性较少	属性较多	属性较多	可多可少
规则长度	较短	不一定	较长	可长可短

8.5 R 语言与粗糙集理论

本节说明如何使用 R 语言中的 **RoughSets**（Riza *et al.*，2014）扩充套件以执行粗糙集理论分析，并以一个简单的人员雇用数据集（Komorowski *et al.*，1999）为例产生粗糙集理论中的各项元素，包含决策表、等价关系、近似空间、简化与规则推演（rule induction）。此人员雇用数据集已内建在 **RoughSets** 扩充套件中，共包含 5 个属性与 8 笔数据，其中，前四项属性为条件属性，第五项则为决策属性，且所有属性均为类别尺度，如表 8.4 所示。

表 8.4 人员雇用数据集

No.	Diploma	Experience	French	Reference	Decision
1	MBA	Medium	Yes	Excellent	Accept
2	MSc	High	Yes	Neutral	Accept
3	MSc	High	Yes	Excellent	Accept
4	MBA	High	No	Good	Accept

续表

No.	Diploma	Experience	French	Reference	Decision
5	MBA	Low	Yes	Neutral	Reject
6	MCE	Low	Yes	Good	Reject
7	MSc	Medium	Yes	Neutral	Reject
8	MCE	Low	No	Excellent	Reject

8.5.1　决策表与等价关系

载入扩充套件与数据集之后，通过 **SF. asDecisionTable** 函数将人员雇用数据集转换为决策表，所有条件属性均为类别型，且第五个属性 Decision 为决策属性。

```
library(RoughSets)
data(RoughSetData)
decision_table<-SF.asDecisionTable(dataset= RoughSetData$hiring,
decision.attr=5,indx.nominal=1:5)
#dataset 自变量为要转换成决策表的数据集
#decision.attr 自变量为指定数据集中的决策属性字段
#indx.nominal 自变量为指定数据集中哪些字段为类别尺度
IND<-BC.IND.relation.RST(decision_table,c(2,3));summary(IND)
```

此外，通过 **BC. IND. relation. RST** 函数可以对决策表中任意条件属性集合产生等价关系结果。例如，指定第二个属性 Experience 与第三个属性 French 可产生以下等价关系对象集合：

$$I_D(\{\text{Experience},\text{French}\}) = \{\{x_1,x_7\},\{x_2,x_3\},\{x_4\},\{x_5,x_6\},\{x_8\}\}$$

8.5.2　近似空间

通过 **BC. LU. approximation. RST** 函数可对决策表在给定的等价关系与决策属性下进一步产生近似空间上下界。以下程序为利用 $D=\{\text{Experience, French}\}$条件属性集合产生的等价关系下与决策属性 Decision 值为 Accept 的对象集合 X 下产生近似空间上下界与界线集合：

$$\underline{D}(X)=\{x_2,x_3,x_4\};\overline{D}(X)=\{x_1,x_2,x_3,x_4,x_7\}$$
$$BN_D(X)=\{x_1,x_7\};pos_D(X)=\underline{D}(X)=\{x_2,x_3,x_4\}$$

```
roughset<-BC.LU.approximation.RST(decision_table, IND)
DX_lower=roughset$lower.approximation$Accept
DX_upper=roughset$upper.approximation$Accept
BN_D=setdiff(DX_upper, DX_lower)
```

此外，给定 $F=\{X_1,X_2\}$，$X_1=\{x_1,x_2,x_3,x_4\}$、$X_2=\{x_5,x_6,x_7,x_8\}$通过以下程序可进一步计算近似集合准确率、分类准确率与近似集合分类质量如下：

$$\alpha_D(X_1)=\frac{3}{5},\quad \alpha_D(X_2)=\frac{3}{5},\quad \alpha_D(F)=\frac{3+3}{5+5}=\frac{6}{10},\quad \gamma_D(F)=\frac{3+3}{8}=\frac{6}{8}$$

```
alpha_D_X=nrow(data.frame(DX_lower))/nrow(data.frame(DX_upper))
DX_lower0=roughset$lower.approximation$Reject
DX_upper0=roughset$upper.approximation$Reject
alpha_D_F=(nrow(data.frame(DX_lower))+nrow(data.frame(DX_lower0)))/
   (nrow(data.frame(DX_upper))+nrow(data.frame(DX_upper0)))
gamma_D_F=(nrow(data.frame(DX_lower))+nrow(data.frame(DX_lower0)))/
   nrow(decision_table)
```

8.5.3 简化与规则推演

通过 **FS.all.reducts.computation** 函数可对指定的决策表产生所有的简化属性集合，借以获知决策表中哪些属性为重要属性。以此人员雇用数据集为例，可通过以下程序得到有两组简化属性集合，分别为{Diploma，Experience}以及{Experience，Reference}，而这两个简化属性集合的交集{Experience}则称为核(core)。

```
res=BC.discernibility.mat.RST(decision_table)
reduct=FS.all.reducts.computation(res);reduct
```

由于此产生 reduct 的程序并未建立在扩充套件中，本节另外以 **RST.rule.induction** 函数建立产生 reduct 的程序，详细程序请见附录程序，同时对每条产生的规则计算支持度、置信度与增益等指标。此外，帕夫拉克(Pawlak，1991)提出的 reduct 产生程序所产生的规则置信度必为 1，但也可能造成支持度过低的情况。因此，在 **RST.rule.induction** 函数中加入设定最小支持度与最小置信度以取得更多潜在规则。以人员雇用数据集为例，若设定最小支持度为 0.25(至少 2 笔数据)所产生的规则共 17 条，其中有两条规则增益未大于 1 将之删除，其余规则如表 8.5 所示。结果显示前两名的规则三个指针均最高，且其条件属性只有 Experience，与前述简化属性集合的核相同，显示 Experience 为关键属性。

```
rule_rst=RST.rule.induction(dataset=RoughSetData$hiring, decision.attr=5,
indx.nominal=1:5, min.sup=0.25); rule_rst
```

表 8.5 人员雇用数据集 RST 规则

No.	Diploma	Experience	French	Reference	Decision	Support	Conf.	Lift
1	—	High	—	—	Accept	0.375	1	2
2	—	Low	—	—	Reject	0.375	1	2
3	—	—	Yes	Excellent	Accept	0.25	1	2
4	MSc	High	—	—	Accept	0.25	1	2
5	—	High	Yes	—	Accept	0.25	1	2
6	MSc	High	Yes	—	Accept	0.25	1	2
7	—	Low	Yes	—	Reject	0.25	1	2
8	MCE	—	—	—	Reject	0.25	1	2
9	MCE	Low	—	—	Reject	0.25	1	2

续表

No.	Diploma	Experience	French	Reference	Decision	Support	Conf.	Lift
10	MBA	—	—	—	Accept	0.25	0.67	1.33
11	—	—	—	Excellent	Accept	0.25	0.67	1.33
12	MSc	—	—	—	Accept	0.25	0.67	1.33
13	MSc	—	Yes	—	Accept	0.25	0.67	1.33
14	—	—	—	Neutral	Reject	0.25	0.67	1.33
15	—	—	Yes	Neutral	Reject	0.25	0.67	1.33

8.6　应用实例——TFT-LCD数组事故诊断

8.6.1　案例简介

本节以TFT-LCD数组事故诊断为例，说明应用粗糙集理论以对大量制程资料进行探索和分析，而缩小工程师事故原因排除的范围，有效率地提供工程师诊断事故原因（Hsu，*et al.*，2010）。在TFT-LCD复杂的制造程序中，不论在数组（array）制程、组立（cell）制程、模块（module）等制程，都会自动化搜集产品通过机台的参数数据，或是以人工方式做记录的判断数据来进行制程监控或故障分析。然而，工程师往往仅由本身的专业知识或经验法则，来分析可能隐藏的异常原因或是归纳出产品质量不良的特征，因此可能受到人为判断和经验而影响决策质量。

8.6.2　分析过程

1. 数据准备

某TFT-LCD厂数组制程共2212片玻璃基板（plate）的制程历史数据，以及各玻璃基板所切割的面板（panel）的缺陷种类及数量，其中每片玻璃基板被切割成15片面板。首先，由制程数据的呈现与检查中，发现有遗漏或错误的数据共101笔，因为占全部数据的比例不大，与领域专家讨论后，决定予以删除。最后整理2111笔观测值，每片玻璃基版的缺陷数则转换为良率，并与领域专家讨论所需要的参数因子多寡、数据的形态、数据格式上的转换后，计算出各玻璃基板的缺陷比率，并将2111片面版良率数据转换为106批次，各玻璃基板在11道制程下所经过的机台，整理后的部分制程分析数据如表8.6所示。并利用K-means分群法将106批次分成高良率（决策属性为1）与低良率（决策属性为0）两组，分群结果如表8.7所列。

表8.6　原始数据包含处理过程的机器和缺陷程度（部分）

批次编号	制程a	完成时间	制程b	完成时间	制程c	完成时间	…	批次良率
AX1PJ02	a04	09/07 20:07:48	b02	09/09 04:49:15	c02	09/10 02:18:24	…	63.75%
AX1PJ03	a04	09/07 19:44:48	b02	09/09 03:54:42	c02	09/10 05:17:02	…	70.37%

续表

批次编号	制程 a	完成时间	制程 b	完成时间	制程 c	完成时间	…	批次良率
AX2PJ01	a03	09/08 21:05:50	b01	09/09 18:37:58	c01	09/13 08:18:23	…	97.00%
AX2PJ02	a01	09/08 22:15:26	b01	09/09 19:28:42	c01	09/11 22:37:01	…	78.33%
AX2PJ04	a02	09/09 03:25:03	b05	09/09 23:35:04	c05	09/13 07:01:33	…	90.33%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
AX2PX01	a01	09/11 02:28:42	b05	09/12 02:15:04	c02	09/15 13:21:23	…	99.67%

表 8.7　两群体的基本统计数据

组　别	数　量	平均值	标准偏差	最小值	最大值	决策属性
高良率	64	92.40%	4.32%	82.33%	100.00%	0
低良率	42	70.68%	7.32%	45.67%	81.05%	1

2. RST 规则提取

根据前面各节所述的粗糙集理论方法和步骤来简化规则。首先，建立一个决策表，如果候选规则的支持度大于 10 则会被接受，其中包含 11 道制程和全部产出量。由于所有的预测属性都需分类，有 384 个简化值没有直接使用离散的训练组数据，根据支持门槛，可以推导出 39 个候选规则，如表 8.8 所示。规则 1 是简化后具有两种属性和支持度 10 的数据，代表“如果一以机器 A03 进行层级 1 的薄膜过程和机器 H03 进行层级 3 的蚀刻过程，则这块薄板被分类为低良率”；规则 2 是另一种简化后具有两种属性和支持度 10 的数据，代表“如果该以机器 B02 进行层级 1 的蚀刻过程、用机器 C06 进行层级 1 的清除过程，则这块薄板为高良率”经过其他 4 次的交叉验证和初步筛选，候选规则 44、40、48 和 39 被分为低良率。此外，个别候选规则均需计算出个别规则的置信度及增益。

以表 8.8 为例，说明制程与过站机台集合 D 与决策集合 F 的相依程度，假设有 65 笔数据，$U=\{x_1,x_2,\cdots,x_{65}\}$，$D=\{a,b,\cdots,k\}$，则 $U|D=\{\{x_1,\cdots,x_{10}\},\{x_{11},\cdots,x_{24}\},\{x_{25},\cdots,x_{37}\},\{x_{38},\cdots,x_{55}\},\{x_{56},\cdots,x_{65}\}\}$，决策集合 $F=\{X_1:\{x_1,x_2,\cdots,x_{37}\},X_2:\{x_{38},x_{39},\cdots,x_{65}\}\}$，集合 X_1 表示为低良率，集合 X_2 表示为高良率。属性集合 D 与决策集合 F 集合 X_1 的下近似为 $\underline{D}(X_1)=\{x_1,x_2,\cdots,x_{37}\}$，属性集合 D 与决策集合 F 集合 X_2 的下近似为 $\underline{D}(X_2)=\{x_{38},x_{39},\cdots,x_{65}\}$，属性集合 D 与决策集合 F 集合 X_1 的上近似为 $\overline{D}(X_1)=\{x_1,x_2,\cdots,x_{37}\}$，集合 X_2 的上近似为 $\overline{D}(X_2)=\{x_{38},x_{39},\cdots,x_{65}\}$，所以此组近似集合分类的准确率 $\alpha_D(F)$，以及此组近似集合分类的质量 $\gamma_D(F)$ 可分别求得为 1，表示通过该属性集合可完全定义决策属性 F。

$$\alpha_D(F)=\frac{\sum card\ \underline{D}(X_i)}{\sum card\overline{D}(X_i)}=\frac{37+28}{37+28}=1$$

$$\gamma_D(F)=\frac{\sum card\ \underline{D}(X_i)}{cardU}=\frac{37+28}{65}=1$$

表 8.8　良率的候选规则(部分)

候选规则	IF(制程 & 过站机台)											THEN(良率)	支持度
	a	b	c	d	e	f	g	h	i	j	k		
1	3	—	—	—	—	—	—	3	—	—	—	低	10
2	—	—	—	—	—	—	—	—	1	—	—	低	14
3	—	—	—	—	2	—	—	3	—	2	—	低	13
4	—	—	—	—	—	—	—	1	—	—	—	高	18
5	—	4	—	—	—	—	—	2	—	—	—	高	10

3. RST 提取规则验证

筛选候选规则的置信度门槛为 70%、增益大于 1,部分候选规则的验证结果如表 8.9 所示。根据第一次交叉验证的结果,规则 1 的置信度是 100%(大于 70%)、增益是 1.75(大于 1),因此会被接受,但规则 2 则会被拒绝,即使其增益是 1.16。

表 8.9　低良率经过第一次交叉验证所产生候选规则的信度

候选规则	规 则 形 式	满足假设条件的样本数	满足假设条件与决策结果的样本数	置信度	增益	规则接受与否
1	若 以机器 a03 进行层级 1 的薄膜制程、用机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率	3	3	100.00%	1.75	Yes
2	若 以机器 b04 进行层级 1 的蚀刻制程和用机器 c06 进行层级 1 的清除制程 则 该玻璃基板为高良率	6	3	50.00%	1.16	No
3	若 以机器 i01 进行层级 3 的清除制程 则 此该玻璃基板为低良率	7	4	57.14%	1	No
4	若 以机器 h01 进行层级 3 的蚀刻制程 则 该玻璃基板为高良率	3	3	100.00%	2.33	Yes
5	若 以机器 e02 进行层级 2 的蚀刻制程和用机器 h03 进行层级 3 的蚀刻制程再用机器 j02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率	1	1	100.00%	1.75	Yes

类似的规则验证过程会持续到所有训练组里的候选规则都被筛选,只剩 13 条规则为止。经过五次的交叉验证和整合筛选出的规则后,可以选出与置信度和增益有关的 18 条规则如表 8.10 所示。

表 8.10　验证后的候选规则

规则	规 则 形 式	置信度	增益
1	若 以机器 a03 进行层级 1 的薄膜制程和机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率(5/5)	87.00%	2.25
2	若 以机器 b02 进行层级 1 的蚀刻制程和机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率(4/5)	96.00%	2.56

续表

规则	规 则 形 式	置信度	增益
3	若 以机器 b02 进行层级 1 的蚀刻制程和机器 h03 进行层级 3 的蚀刻制程和机器 j02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率(5/5)	91.50%	2.47
4	若 以机器 b02 进行层级 1 的蚀刻制程和机器 j02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率(5/5)	76.00%	1.97
5	若 以机器 e02 进行层级 2 的蚀刻制程和机器 h03 进行层级 3 的蚀刻制程再用机器 j02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率(5/5)	88.40%	2.31
6	若 以机器 e02 进行层级 2 蚀刻制程和机器 j02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率(5/5)	81.40%	2.16
7	若 以机器 e03 进行层级 2 的蚀刻制程 则 该玻璃基板为低良率(1/5)	100.00%	1.75
8	若 以机器 f03 进行层级 2 的清除制程和机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率(5/5)	83.40%	2.14
9	若 以机器 h03 进行层级 3 的蚀刻制程和机器 i01 进行层级 3 的清除制程 则 该玻璃基板为低良率(5/5)	85.00%	2.15
10	若 以机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率(4/5)	79.33%	1.95
11	若 以机器 j02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率(2/5)	88.89%	1.85
12	若 以机器 a06 进行层级 1 的薄膜制程和机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率(4/5)	81.25%	2.29
13	若 以机器 a03 进行层级 1 的薄膜制程 则 该玻璃基板为低良率(4/5)	86.67%	2.29
14	若 以机器 e02 进行层级 2 的蚀刻制程,机器 h03 进行层级 3 的蚀刻制程和机器 I01 进行层级 3 的清除制程 则 该玻璃基板为低良率(1/5)	100.00%	2.33
15	若 以机器 h03 进行层级 3 的蚀刻制程,机器 i01 进行层级 3 的清除制程再用机器 J02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率(2/5)	76.67%	2.16
16	若 以机器 h03 进行层级 3 的蚀刻制程和机器 j02 进行层级 4 的薄膜制程 则 该玻璃基板为低良率(4/5)	83.67%	2.42
17	若 以机器 e03 进行层级 2 的蚀刻制程和机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率(2/5)	70.00%	1.98
18	若 以机器 e02 进行层级 2 的蚀刻制程和机器 h03 进行层级 3 的蚀刻制程 则 该玻璃基板为低良率(2/5)	81.50%	2.35

8.6.3 案例小结

本案例以某 TFT-LCD 厂数据为实证,以检验本研究架构之效度。根据实证结果发现,粗糙集理论能提出有用的规则,协助工程师缩小事故发生原因的搜索范围,找出问题的根源并能提供信息帮助决策者解决问题。未来应针对各种不同的实证方法做进一步研究,找出

TFT-LCD 在不同阶段之复杂的制程间的相互关系,进而开发出更好的分析方法以提高效率及产量。

8.7　结论

粗糙集理论可以用来归约数据集合、简化属性、挖掘隐藏在数据中的样型,并从数据中产生最小集合的决策规则,而且能够直观地解释所获得的结果,因此可以作为推导分类或决策规则的数据挖掘方法(Kusiak, 2001; Walczak & Massart, 1999;Pawlak, 1982)。

粗糙集理论最大的限制是在处理属性为连续型变量时,必须将数据离散化(Pawlak, 1997, 1996, 1982)。由于连续型数据变量的可能值会出现在一个范围,而数据点太多将使结果过于分散,可能造成该属性不容易产生规则,因此如要使规则有较高的准确率,必须将连续型数据转换成离散型的数据。

问题与讨论

1. 假设定义三个属性与各元素:教育程度 D_1、职业 D_2、性别 D_3,

$$U=\{\text{Allen},\text{Bob},\text{Carl},\text{Dennis},\text{Eva},\text{Frank},\text{Grace},\text{Helen},\text{Ivy},\text{Jason}\}$$
$$=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$$

根据以上属性定义以及下表建立近似空间 $U|D$。

关系 D 及 U

D \ U	教育程度 D_1	职业 D_2	性别 D_3
Allen	大学	老师	女
Bob	研究所	工程师	男
Catherine	研究所	医生	女
Dennis	大学	工程师	男
Eva	大学	医生	女
Frank	研究所	老师	男
Grace	大学	医生	女
Helen	高中	工程师	女
Ivy	大学	医生	女
Jason	研究所	工程师	男

2. 承上题,假设 $X=\{x_3,x_5,x_7\}$,试找出上表中 D_1、D_2、D_3 的上限近似集合与下限近似集合。

3. 承上题,试找出上表中 D_1、D_2、D_3 的近似集合的准确率。

4. 承上题,试求出上表中的 reducts。

5. 请比较粗糙集理论、决策树分析、贝叶斯网络、关联规则方法之间的优缺点。

附录程序(RST. rule. induction)

```
RST.rule.induction<-function (dataset, decision.attr=NULL, indx.nominal=NULL, min.sup=NULL, min.conf
=NULL){
  require(RoughSets)
  decision.table <-SF.asDecisionTable(dataset, decision.attr, indx.nominal)
  n <-nrow(decision.table)
  p <-ncol(decision.table)-1
  pset <-setdiff(seq(1,ncol(decision.table)),decision.attr)

  rule={};support.n={};conf.X={};lift.Y={}
  for (i in 1:n){
    Dset=as.character(which(decision.table[,decision.attr]==decision.table[i,decision.attr]))
    for (m in 1:(p-1)){
      comb=combn(p,m)
      for (j in 1:ncol(comb)) {
       set=decision.table
       for(k in 1:m) set=set[set[,pset[comb[k,j]]]==decision.table[i,pset[comb[k,j]]],]
       Cset=rownames(set)

       if (is.null(min.sup)) {
         if (is.null(min.conf)){
           if (setequal(intersect(Dset, Cset), Cset)){
               reduct=decision.table[i,]
               fe=setdiff(seq(1,ncol(decision.table)),union(decision.attr,pset[comb[,j]]))
               reduct[,fe]<-"x"
               rule=rbind(rule,reduct)
               support.n=c(support.n,length(intersect(Dset, Cset)))
               conf.X=c(conf.X,length(Cset))
               lift.Y=c(lift.Y,length(Dset))
           }
         } else {
          if (length(intersect(Dset, Cset))/length(Cset) > =min.conf) {
             reduct=decision.table[i,]
             fe=setdiff(seq(1,ncol(decision.table)),union(decision.attr,pset[comb[,j]]))
             reduct[,fe]<-"x"
             rule=rbind(rule,reduct)
             support.n=c(support.n,length(intersect(Dset, Cset)))
             conf.X=c(conf.X,length(Cset))
             lift.Y=c(lift.Y,length(Dset))
          }
         }
       } else if (is.null(min.conf)){
          if (length(intersect(Dset, Cset))/n > =min.sup) {
             reduct=decision.table[i,]
             fe=setdiff(seq(1,ncol(decision.table)),union(decision.attr,pset[comb[,j]]))
```

```
                reduct[,fe]<-"x"
                rule=rbind(rule,reduct)
                support.n=c(support.n,length(intersect(Dset, Cset)))
                conf.X=c(conf.X,length(Cset))
                lift.Y=c(lift.Y,length(Dset))
              }
          } else {
              if (length(intersect(Dset, Cset))/n > =min.sup & length(intersect(Dset, Cset))/length(Cset)
              > =min.conf) {
                reduct=decision.table[i,]
                fe=setdiff(seq(1,ncol(decision.table)),union(decision.attr,pset[comb[,j]]))
                reduct[,fe]<-"x"
                rule=rbind(rule,reduct)
                support.n=c(support.n,length(intersect(Dset, Cset)))
                conf.X=c(conf.X,length(Cset))
                lift.Y=c(lift.Y,length(Dset))
              }
          }
         }
       }
    }

    if (! is.null(rule)) {
      rule2=rule[! duplicated(rule),]
      support.n=support.n[! duplicated(rule)]
      conf.X=conf.X[! duplicated(rule)]
      lift.Y=lift.Y[! duplicated(rule)]
      rownames(rule2)<-seq(1,nrow(rule2))
      support=support.n/n
      conf=support.n/conf.X
      lift=conf/(lift.Y/n)
      rule=cbind(rule2,support,conf,lift)
    }
    return(rule)
}
```

第 9 章

预测与时间数据分析

预测是推测未来的过程,常以过去的历史数据(historical data)为依据。例如,预测将来的销售量、股价以及客户消费行为等。**多变量分析(multivariate statistical analysis)**主要用于分析多个变量间的关联、发掘其背后可能存在的样型,根据有无相依变量、不同的数据尺度与变量个数,可以采用不同的分析方法,本章对数据挖掘应用上常见的回归分析与逻辑回归进行说明。其他多变量分析方法可参照(Johnson & Wichern,2007)、(Hair *et al.*,2010)等。**时间序列数据(time series data)**是依据规律时间间距下连续观察的量测值,通过分析已发生的时间序列数据的特性,来预测未来值的过程。

9.1 回归分析

回归分析(regression analysis)是分析一个或多个独立变量(independent variable)对某一个相依变量(dependent variable)的相关程度,也可了解当独立变量改变时,对相依变量的影响(Draper & Smith,1981),例如,经济成长率对手机销售量的影响。独立变量是解释变量或预测变量,而相依变量则是反应变量。

9.1.1 回归分析基本介绍

散布图(scatter diagram)是表示两变量间关系的基本工具。通常 X 轴(横轴)代表独立变量,Y 轴(纵轴)代表相依变量,散布图中的数据代表独立变量与相依变量的成对数据。若数据点分布于狭长的带状区域内,当独立变量值增加时,相依变量值也会依比例增加(正相关)或减少(负相关),则称两变量间具有线性关系(如图 9.1(a))。借由观察散布图中所有数据显示的形状、方向,可初步判断变量间关系强度。若数据点均靠近一条直线,即称两变量间存在高度线性关系(如图 9.1(a));若数据点为不规则的散布,且无线性趋势,则表示变量间不存在线性关系或低度线性相关(如图 9.1(b))。

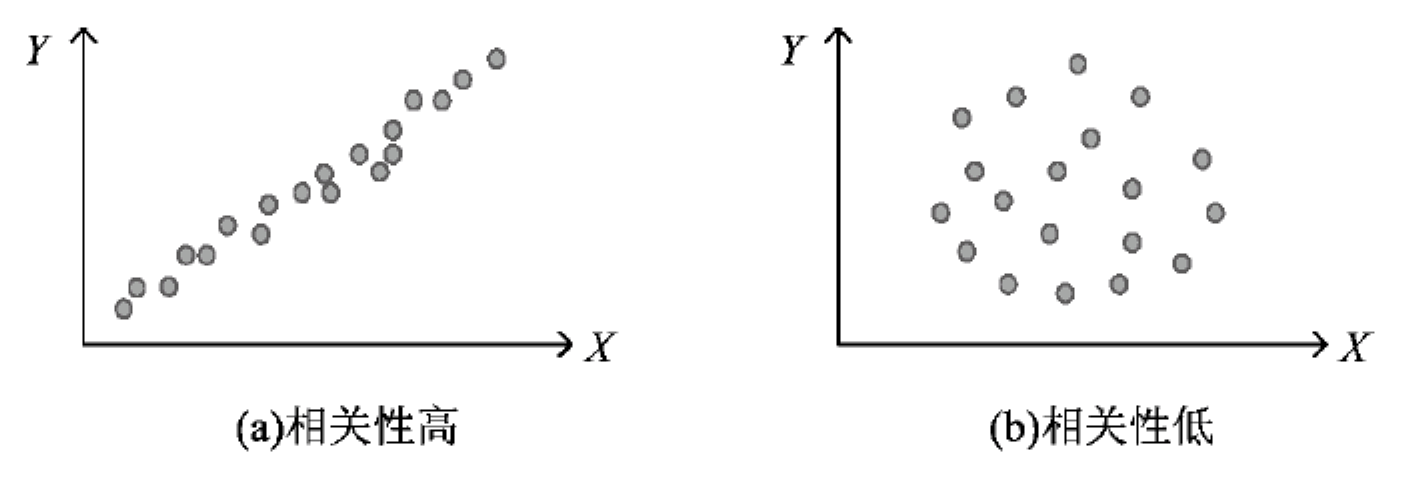


图 9.1　两变数 *X-Y* 的相关程度

散布图可提供配适回归线形态的参考依据,可借由描绘独立变量与相依变量间的散布情形,决定合适的回归函数形态,以进行模式建立、估计以及预测。当变量的关系为曲线相关或非线性相关时,可能无法仅用线性函数来配适其模型,如图 9.2(a),而需用非线性函数来描述变量的关系,如图 9.2(b)。

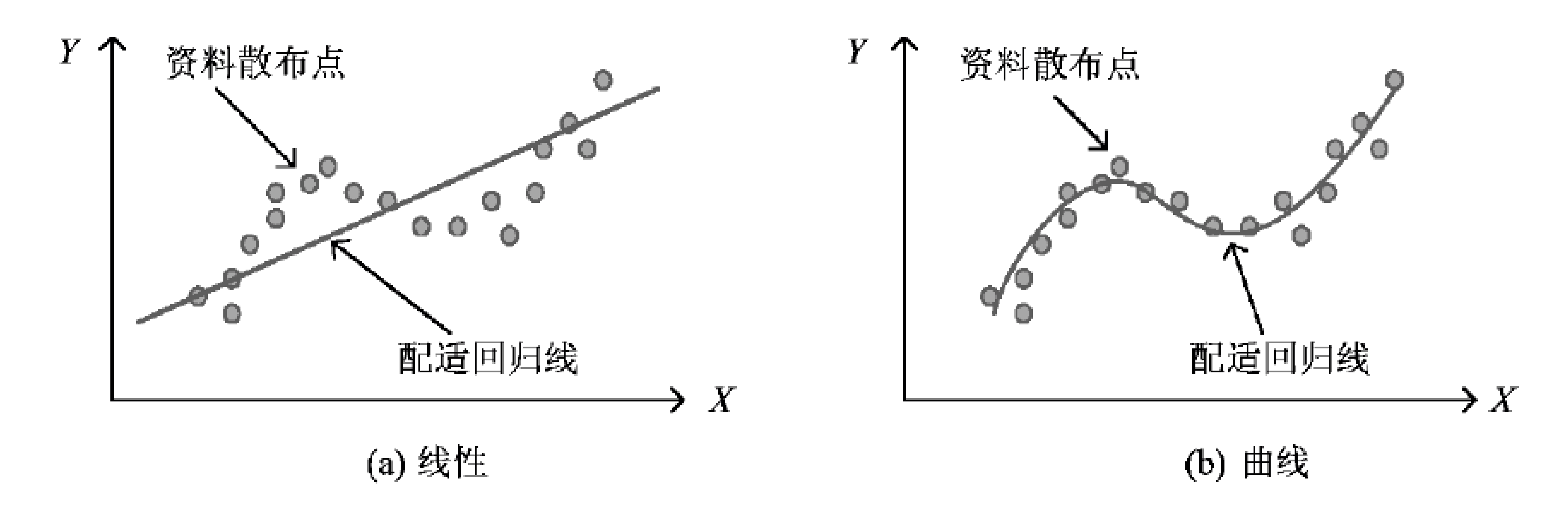


图 9.2　线型及曲线型配适函数

散布图亦可作为筛选独立变量的基本方法,如两变量的关系度很低,表示该独立变量对相依变量的解释能力有限,可以从模式中剔除该独立变量。数据的相关系数仅能显示两变量间是否存在相关,无法确定哪一个是因、哪一个是果,甚至可能完全不存在因果关系。例如,搜集 200 位男性上班族的收入与体脂肪数据,得到结果发现,收入越高则其体脂肪越高,所以降低收入是否即可降低体脂肪?实际上收入与体脂肪都受到年龄的影响,因为年龄越大,受到代谢降低的影响,所以体脂肪自然容易上升;另一方面,年龄越大表示工作服务的年资越长,所以收入一般而言也会比较高,使得表面上看起来收入与体脂肪有相关。

回归分析是建立一个或多个独立变量对某一相依变量的关系模式,借由回归方程式中参数的估计,可以评估独立变量对相依变量的贡献或影响程度。回归分析可分为**单回归**(**simple regression**)与**多重回归**(**multiple regression**)。单回归是描述一个独立变量对一个相依变量的关系;多重回归则用以描述多个独立变量对一个相依变量的关系。

[**范例 9.1**]　假设公司销售业绩与公司营收间存在关联,若公司营销经理想预测公司未来营收的走向,若你是业务人员是否能利用销售业绩来提供未来公司营收的预测。如搜集过去 10 年某公司产能、产品平均售价与公司的总营收数据如表 9.1。若欲了解产品平均售价与总营收之间的关系,以单回归为例,则可绘其散布图如图 9.3。从图 9.3 中可发现产品平均售价与总营收呈现正相关,其单线性回归模式为 $\hat{Y}=-0.978+3.244x$,可作为下一年度的预测模型。如下一年的产品平均售价为 1.85 万元时,则预测下一年公司总营收约为 500 万元。

表 9.1　某公司近十年的产能利用率、产品平均售价与总营收数据

年　　份	1	2	3	4	5	6	7	8	9	10
产品平均售价 x/万元	1.8	1.6	1.9	1.7	1.8	2.0	2.1	2.2	2.0	2.1
总营收 $\hat{Y}$/百万元	5.1	3.9	4.5	5.2	4.1	5.8	6.2	5.4	6.3	6.0

公司在法人说明会提出未来营收预测,例如,可根据过去几年产品平均售价与公司营收的数据做下一年的预测(可采用单回归分析);亦可根据产能利用率、产品组合、技术组合及

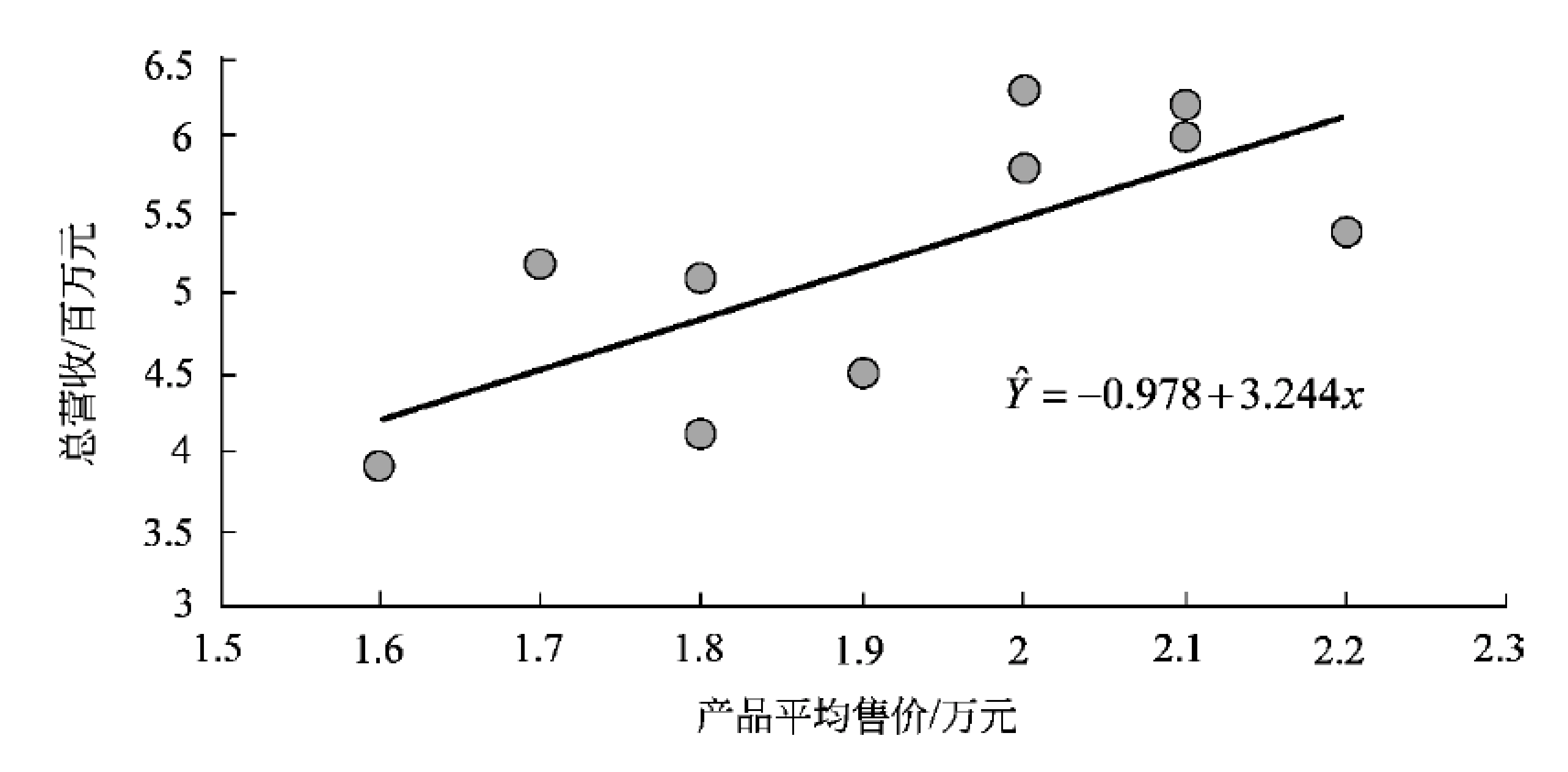


图 9.3 过去 10 年总营收 $\hat{Y}$ 对产品平均售价 x 的散布图与回归线

产能配置等做预测(可采用多重回归分析);若营收数据间存在自相关(autocorrelation),则可根据过去几年营收数据做下年度的预测(可采用时间数据分析)。

如图 9.4 所示,假设在特定的 x_i 值下,经由重复实验所搜集到对应的 y_i 值,而搜集的母体成对数据形成一个概率密度函数 $f(y_i)$(probability density function),且该母体概率密度函数的平均数 $E(Y|x)=\mu_{y|x}$ 落在相依变量 Y 对独立变量 x 的母体回归线 $E(Y|x)=\beta_0+\beta_1 x$ 上,其中,β_0,β_1 为回归系数,β_0 是指母体回归线在纵轴上的截距,也就是回归线与原点之间的距离;β_1 则是回归线的斜率,它表示 x 每增加 1 单位所引起 Y 的增量。

然而,实际量测值也受到其他未被考虑的因素或随机误差影响,可利用误差项 ε_i(error term)以表示在相同实际测量或搜集的 y_i 与母体回归线的平均数间的误差,ε_i 是随机变量,在回归分析中假设 ε_i 服从正态分配。因此,若 x 与 Y 变量间存在线性相关,则相依变量 Y 可用母体回归线与随机误差项来表示,如式(9.1)。

$$Y_i=\mu_{Y|x}+\varepsilon_i=\beta_0+\beta_1 x_i+\varepsilon_i,\quad i=1,2,\cdots,n \tag{9.1}$$

母体回归线可代表两变量间的线性关系,由于母体回归线无法得知,而改以样本回归线 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$ 来估计母体回归线,$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 又称为样本回归系数。图 9.4 说明母体回归线与样本回归线之间的关系,其中,ε_i 为母体回归线的随机误差,e_i 代表实际量测值 Y_i 与样本回归线估计值 $\hat{Y}_i$ 的差距,即 $e_i=y_i-\hat{y}_i$,又称残差(residual)。残差包含了未被考虑的因素所造成的潜在的系统性误差和随机误差。

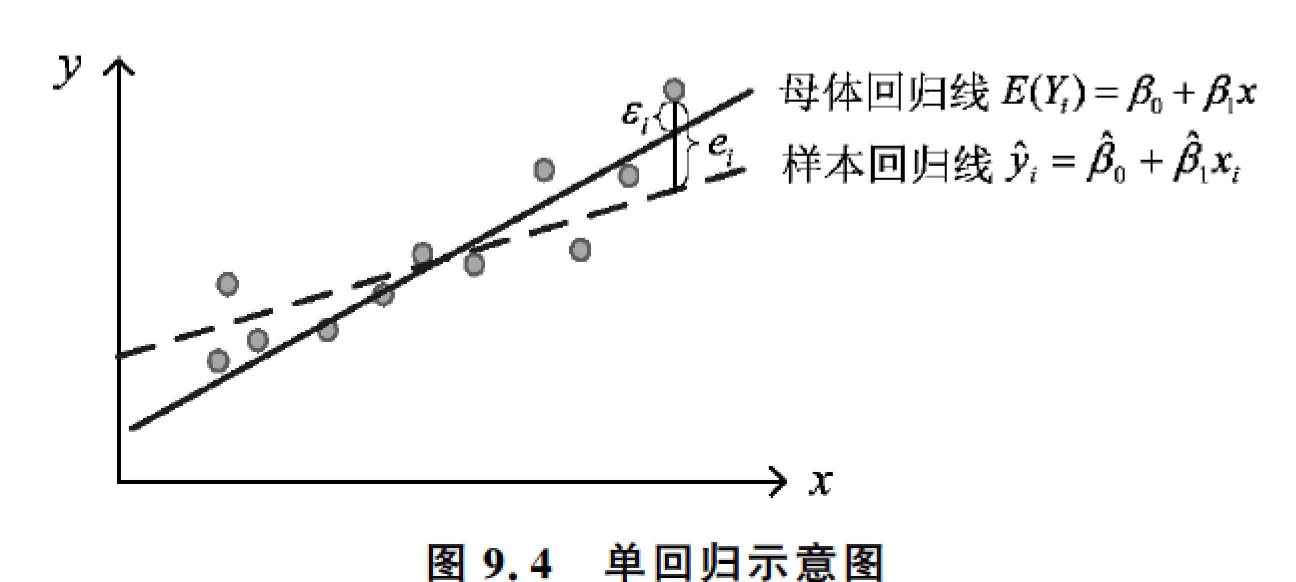


图 9.4 单回归示意图

9.1.2 参数估计

最小二乘估计法(least squares estimate method)是以最小化残差平方和(sum of squared error, SSE),找出最接近母体 Y 的样本回归模型,样本回归式的残差平方和越小,表示以此样本回归模式配适此组数据的结果越好。

$$\text{SSE} = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{9.2}$$

为了求得 SSE 最小,可分别对 $\hat{\beta}_0$,$\hat{\beta}_1$ 偏微分。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} \tag{9.3}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x} \tag{9.4}$$

其中,$\bar{x}$ 为独立变量样本平均数,$\bar{Y}$ 为相依变量样本平均数,由于母体回归线的真实方差 σ^2 未知,因此,用其不偏估计量 $\hat{\sigma}^2$ 来估计 σ^2,$\hat{\sigma}^2$ 又称均方误差(mean squared error,MSE),其公式如下:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2} = \text{MSE} \tag{9.5}$$

9.1.3 回归模型解释与评估

回归分析常用于预测,借由数据库中某些已知的信息以预测未知的变量。如果独立变量之间存在**共线性(collinearity)**,则容易发生模型解释能力高,但个别变量检定不显著的问题。若想了解回归模式的解释能力,可以利用独立变量来预测相依变量的能力,其相关性是否具统计上显著的意义?哪些独立变量对相依变量比较重要,说明如下。

回归模型的**拟合优度(goodness of fit)**检定,可以比较加入独立变量 x 的信息后,对于解释或预测相依变量 y 的能力提升多少,作为回归模型拟合优度的衡量,并可借由回归模型将相依变量的总平方和(total sum of squares, SST)分解为可解释的平方和,又称为回归平方和(sum of squares due to regression, SSR)与不可解释的平方和,又称为残差平方和(sum of squares due to error, SSE),总平方和即为回归平方和与残差平方和之和,即 SST=SSR+SSE。以图 9.5 为例说明其关系,$(\bar{x},\bar{y})$为样本平均值,(x_i,y_i)为某样本数据,$(x_i,\hat{y}_i)$为该样本所对应的回归估计值。

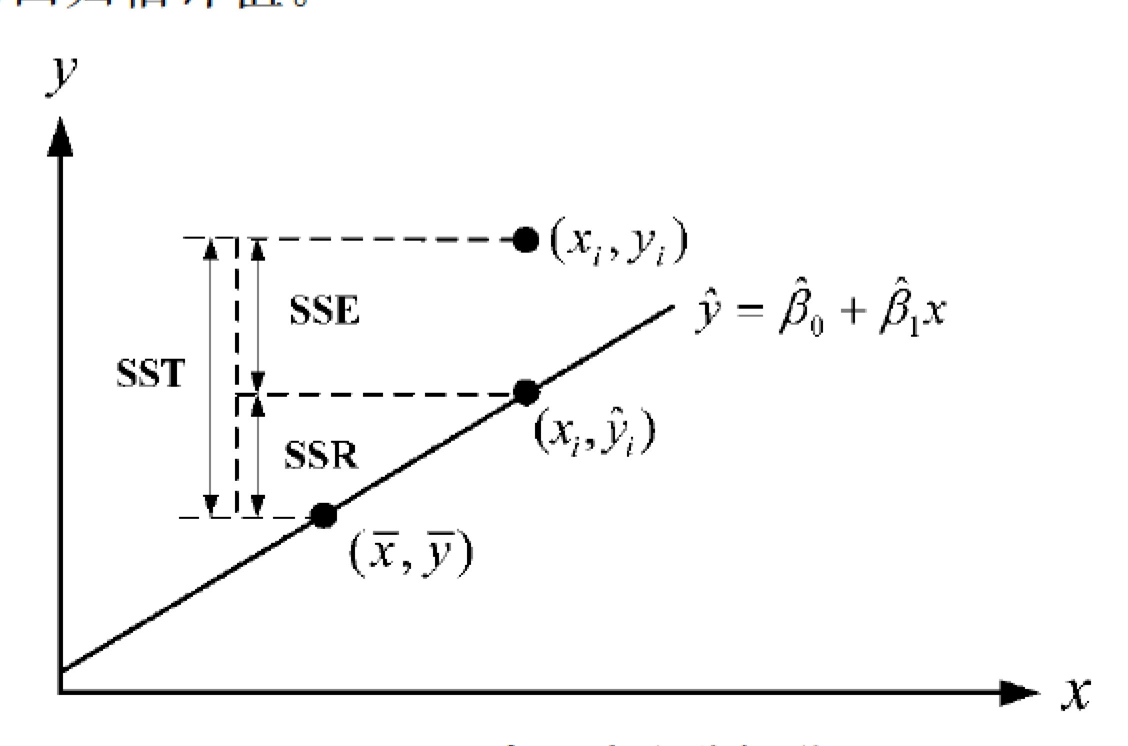


图 9.5　回归平方和分解说明

因此，可根据可解释的平方和占总平方和的比例作为衡量该回归模型的拟合优度，当 SSR/SST 越大，表示总平方和中有越多的比例可被该回归模型解释，也说明回归模型的拟合优度佳。回归之方差分析其假设检定为 $H_0:\beta_1=0, H_1:\beta_1\neq 0$

表 9.2 单回归的方差分析表

变异来源	平方和	自由度	均方和	检定统计量
回归模型	$SSR=\sum_i(\hat{y}_i-\bar{y})^2$	1	$MSR=SSR/1$	$F=\frac{MSR}{MSE}$
残差	$SSE=\sum_i(y_i-\hat{y})^2$	$n-2$	$MSE=SSE/(n-2)$	
总和	$SST=\sum_i(y_i-\bar{y})^2$	$n-1$		

若检定统计量 $F>F_{(1-\alpha,1,n-2)}$，则拒绝虚无假设，表示此回归模型系数显著不为 0。以[范例 9.1]为例，其方差分析结果如表 9.3 所示。

表 9.3 售价与营收单回归模式的方差分析表

变异来源	平方和	自由度	均方和	检定统计量
回归模型	3.536	1.000	3.536	9.158
残差	3.089	8.000	0.386	
总和	6.625			

在显著水平 $\alpha=0.05$ 下，检定结果 $F=9.158>F_{(0.95,1,8)}=5.318$，因此拒绝虚无假设，显示该回归模型显著。

除了检查整体回归模型拟合优度外，也可通过检定个别回归系数是否显著不为 0，若该独立变量对于相依变量有解释能力，则其 t 检定判定结果应为显著。简单回归模型中回归系数 $\hat{\beta}_1$ 的假设检定说明如下：

$$H_0:\beta_1=\beta^*,\quad H_1:\beta_1\neq\beta^*\text{（一般而言通常假设 }\beta^*=0\text{）}$$

在 H_0 为真的情况下，

$$\hat{\beta}_1\sim N\left(\beta_1,\frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}\right)$$

$$\text{故检定统计量 } t=\left|\frac{\hat{\beta}_1-\beta_1}{\frac{\hat{\sigma}^2\sqrt{n-1}}{\sum(x_i-\bar{x})^2}}\right|$$

若 $t>t_{(1-\alpha/2,n-1)}$ 则拒绝虚无假设，表示回归系数不显著。回归系数不显著的原因，可能是独立变量与相依变量之间无线性相关，也有可能是独立变量间的共线性所造成。

模型配适后，经由判定系数(determinant of coefficient)R^2 来判断及衡量所构建模式的解释能力，如式(9.6)：

$$R^2=1-\frac{SSE}{SST}=\frac{SSR}{SST}\tag{9.6}$$

R^2 表示在考虑所有独立变量下，解释相依变量的平方和百分比或其预测解释能力，可以用来代表线性回归模式的拟合优度。然而，R^2 值越大，并不一定表示回归模型配适得越好，因为只要独立变量的个数增加，模式的 R^2 即会增加，造成过度配适现象。因此，常使用调整后判定系数(adjusted determinant of coefficient)R_a^2 取代 R^2 以作为模式评估的基准，如式(9.7)所示：

$$R_a^2 = 1-(1-R^2)\cdot\frac{n-1}{n-p} = 1-\frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1-\frac{n-1}{\text{SST}}\cdot\text{MSE} \tag{9.7}$$

其中，n 为样本观测值个数，而 p 为样本回归模型中所选取的参数个数。

9.1.4　多重回归分析

当相依变量受到多个独立变量影响时，可利用多重回归分析了解各个独立变量的影响。例如式(9.8)为多重回归模型：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{9.8}$$

其中，β_0，β_1，β_2 都是母体参数，参数 β_1 是当 x_2 固定时，x_1 每增加一单位所引起的平均数对应值增量的变动；同样地，参数 β_2 是当 x_1 固定时，x_2 每增加一单位所引起的平均数对应值增量的变动。倘若 x_1 对平均数对应值的影响不依赖 x_2 的水平或 x_2 对平均数对应值的影响不依赖 x_1 的水平，则此两变量称为无交互作用。

β_0，β_1，β_2 都是参数且为未知数，称为偏回归系数(partial regression coefficient)，因为模式中的某一独立变量是固定数量而对相依变量的影响只是来自另一个变量的变动。也可以微积分方式诠释回归系数意涵，也就是分别对式(9.8)求 x_1 与 x_2 的偏微分：

$$\left.\begin{aligned}\partial E(Y)/\partial x_1 &= \beta_1\\ \partial E(Y)/\partial x_2 &= \beta_2\end{aligned}\right\} \tag{9.9}$$

即当某一独立变量为固定数量时，另一个独立变量每单位变动所引起 $E(Y)$ 变动的比例。例如，$E(Y)=15+2x_1-3x_2$，欲从两变量的变动计算期望值 $E(Y)$，则当 x_2 为固定数量，每单位 x_1 的变动，期望值 $E(Y)$ 将随之增加 2 倍；当 x_1 为固定数量，每单位 x_2 的变动，期望值 $E(Y)$ 将随之减少 3 倍。

9.1.5　共线性

当独立变量间存在高度相关性，称为共线性或多重共线性(multicollinearity)，可能导致回归方程式显著，但各自独立变量的回归系数估计偏差或不显著，使得回归分析结果难以解释。因此，应尽量消除共线性对数据分析和建模的影响。

共线性的检定方式可经由方差膨胀因子(variance inflation factor，VIF)的大小来衡量，衡量某一变量与其他变量是否相关的方式为将该独立变量视为其他变量的相依变量，定义 R_i^2 代表该独立变量可被其他独立变量解释变异的比例，则可定义容忍度(tolerance)为 $1-R_i^2$，代表该独立变量无法被其他独立变量解释的残差大小，所以若 R_i^2 越大，则其容忍度越小，VIF 值越大，表示该变量无法被解释的残差比例越低，共线性的程度越明显。实际上，VIF 值为容忍度的倒数，$\text{VIF}=1/\text{tolerance}=1/(1-R_i^2)$。一般而言，若容忍度小于 0.1 代表存在高度共线性问题。

处理共线性的方法除可利用逐步选取(stepwise)的方式外，也可以采用主成分分析将

数据转换为数个直交的主成分。

9.2 逻辑回归

逻辑回归(logistic regression analysis)是处理相依变量为类别变量、独立变量为连续变量的方法。

9.2.1 概率与胜算

表 9.4 为抽烟习惯与有无肺癌的列联表,其中,抽取了 20 位有肺癌的病患以及 180 位无肺癌的病患,经由表 9.4 可得以下概率:

表 9.4 抽烟习惯与有无肺癌的列联表

	有抽烟习惯(*S*)	无抽烟习惯(*NS*)	总 和
肺癌(*H*)	10	10	20
无肺癌(*N*)	30	150	180
总和	40	160	200

得肺癌的概率为 $P(H)=20/200=1/10$;

有抽烟习惯的病患中,得肺癌的概率为 $P(H|S)=10/40=1/4$;

无抽烟习惯的病患中,得肺癌的概率为 $P(N|S)=10/160=1/16$。

在实际应用上,若发生的结果仅有两种,例如发生或不发生,分析者可以用胜算(**odds**)作为分析依据,将该事件发生的概率除以不发生的概率,例如球赛的胜算、赢得大乐透的胜算,以表 9.4 中,得肺癌的胜算为 $\text{odds}(H)=20/180=1/9$,表示所有患者中有无肺癌的概率相同。此外,若有抽烟的病患中有肺癌的胜算为 $\text{odds}(H|S)=10/30=1/3$,表示抽烟的病患中罹患肺癌对没有罹患肺癌的比值是 1 比 3,也可说抽烟者罹患肺癌的胜算是没有罹患肺癌的 1/3 倍。

概率与胜算为不同型式,但均提供相同的信息与结果,彼此间也很容易转换。

$$\text{odds}(H \mid S) = \frac{P(H \mid S)}{P(\bar{H} \mid S)} = \frac{P(H \mid S)}{1 - P(H \mid S)} = \frac{1/4}{3/4} = \frac{1}{3}$$

$$P(H \mid S) = \frac{\text{odds}(H \mid S)}{1 + \text{odds}(H \mid S)} = \frac{1/3}{1 + 1/3} = \frac{1}{4}$$

9.2.2 逻辑回归模式

假设有 k 个独立变量 $x_1, x_2, \cdots, x_k$ 与一二元相依变数(0,1)时,逻辑回归主要用来描述独立变量与相依变量等于 1 的概率。其概率模式如式(9.10)所示,对应概率的值域落在 0 与 1 之间。

$$p = e^{f(x)}/(1 + e^{f(x)}) \tag{9.10}$$

其中,$f(x)$为 x 的多项式函数,若为线性多项式 $f(x)=\beta_0+\beta_1 x+\beta_2 x_2+\cdots+\beta_k x_k$,函数图形如图 9.6 所示。

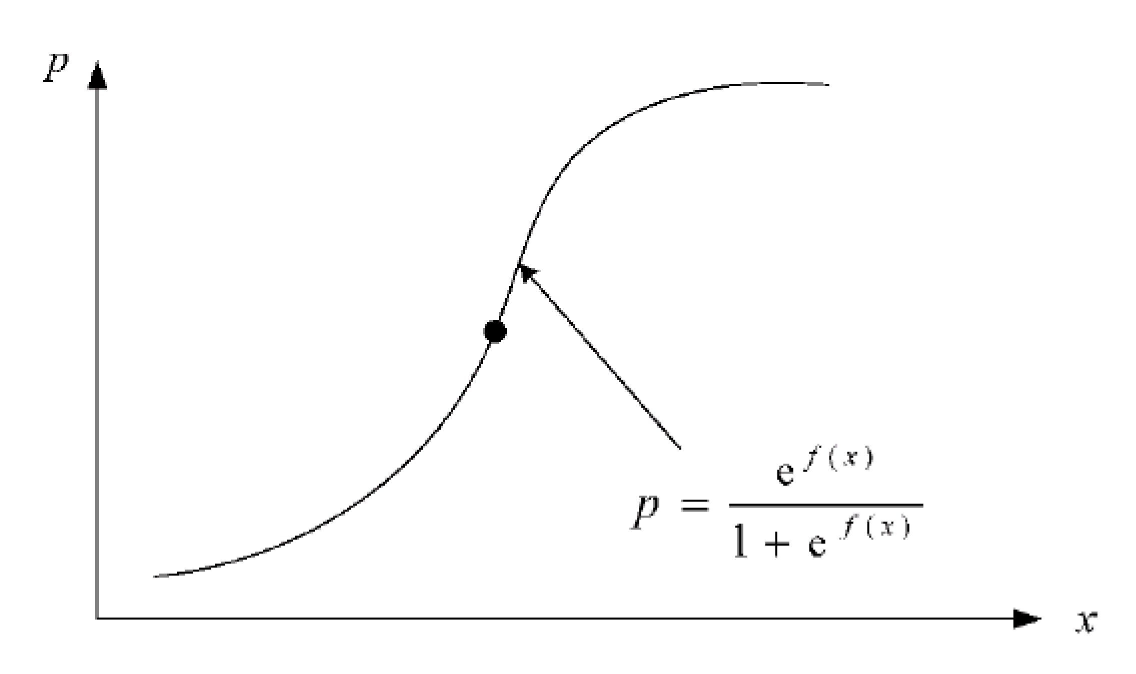


图 9.6　逻辑对应函数形式

逻辑函数为非线性函数，经过适当地转换后，可采用线性模式形态配适数据以良好地描述 p 与 $x_1, x_2, \cdots, x_k$ 的关系，可令 p 表示某种事件成功的概率，受独立变量 $x_1, x_2, \cdots, x_k$ 的影响，若 p 与 $x_1, x_2, \cdots, x_k$ 的关系函数如同式(9.10)，则该事件失败的概率为 $1-p$，如式(9.11)所示。

$$1-p = 1/(1+e^{f(x)}) \tag{9.11}$$

故其胜算为成功的概率对失败概率的比值，如式(9.12)：

$$p/(1-p) = e^{f(x)} \tag{9.12}$$

将式(9.12)取自然对数(ln)后得式(9.13)：

$$\ln[p/(1-p)] = f(x) = \beta_0 + \beta_1 x + \beta_2 x_k + \cdots + \beta_k x_k \tag{9.13}$$

如式(9.13)所示，将胜算取对数后，即可以多重回归分析进行数据配适以及模式构建。

逻辑回归的概率 p 与独立变量间为非线性关系，ln(odds)与独立变量间为线性关系，因此逻辑回归所求得的回归系数是针对 ln(odds)，并非对 p。逻辑回归中模式的参数估计是利用最大似然估计法，相关证明有兴趣的读者可参阅(Sharma，1996)或(Johnson & Wichern，2007)。

逻辑回归模型以“ln(odds)”作为独立变量的线性组合函数，即利用自然对数转换的方式而使逻辑对应函数亦能具有线性性质，以简化非线性函数后续分析的不便以及复杂度，如参数估计、回归系数显著性检定、模式稳健性等。且使用在复回归对每个系数的 t 检定，相当于在逻辑回归中检定每个独立变量的系数是否为 0 的卡方检定。由于逻辑回归模型的反相依变量是以二元指示变数呈现，故与一般回归系数属量检定法不同，应采以属质检定法(如卡方检定)进行系数检定。

以营销顾客细分为例。假设某家寝具用品公司欲制作邮购产品目录以吸引顾客群，寄发给该城市非会员的 200 000 位顾客。其相依变量为“该收件者是否会下单购买产品”，独立变量以五个具代表性的特征表示：X_1 为顾客于过去 3 个月内是否曾购买相关产品；X_2 为该城市的单身人口比例；X_3 为顾客每月所得收入；X_4 为顾客性别；X_5 为顾客居住地为自有或租赁状况。使用逻辑回归以购买的概率 p 作为五个独立变量的函数所配适的模型为

$$\ln[p/(1-p)] = -0.35x_1 - 0.47x_2 + 0.53x_3 + 0.28x_4 + 0.6x_5$$

$p/(1-p)$是成功结果的胜算，系数的正负号表示该独立变量与此一模型反相依变量的关系是正向抑或负向。从上述等式中可以发现成功概率与右式值成正比。而此一模型的预测值允许估计顾客会从此邮购目录购买的概率 p。在未来，此模型可应用于只寄目录给估计会

购买概率超过某一切点的顾客，以使营销资源达到最佳效益。

9.3 时间序列分析

时间序列分析的目的是经由分析时间序列数据的自相关，以及各种形态，如趋势、季节、介入事件等特性，归纳并估计能反映历史数据的时间序列模式。时间序列分析可依照单一变量历史数据的相关性建立模式，并假设单一变量相隔的时间越短，彼此的相关程度就越高。

依观察值属于连续型或离散型，又可分为连续型时间序列与离散型时间序列。时间序列一般呈随机分布，即对序列未来结果无法确定，以概率分配表示，称为未确定时间序列(non-deterministic time series)或随机性时间序列(stochastic time series)；若时间序列是随着数学函数而变化，预测未来的结果为固定的，则为确定性时间序列(deterministic time series)。时间序列数据形态多为随机性时间序列，可依其序列特性及波动情况区分为下列五种形态，如图 9.7 所示。

平稳型时间序列(stationary time series)的观测值是在同一固定水平与固定区域之间变动，且这种特征不随时间变化而改变，如图 9.7(a)所示。在无特殊改变或离群值的情况下，可合理推论此类序列未来的观察值仍在同一水平与区间变动；此外，亦可借由连续观察值间的相依性来提高预测效果。例如，若连续观察值间趋向于负相关，则在得到一正向观测值时，可推测序列的下一观测值为负向观测值的机会居高。若能以一概率函数来配适观察值间的相关性，则能得到有效的预测结果。

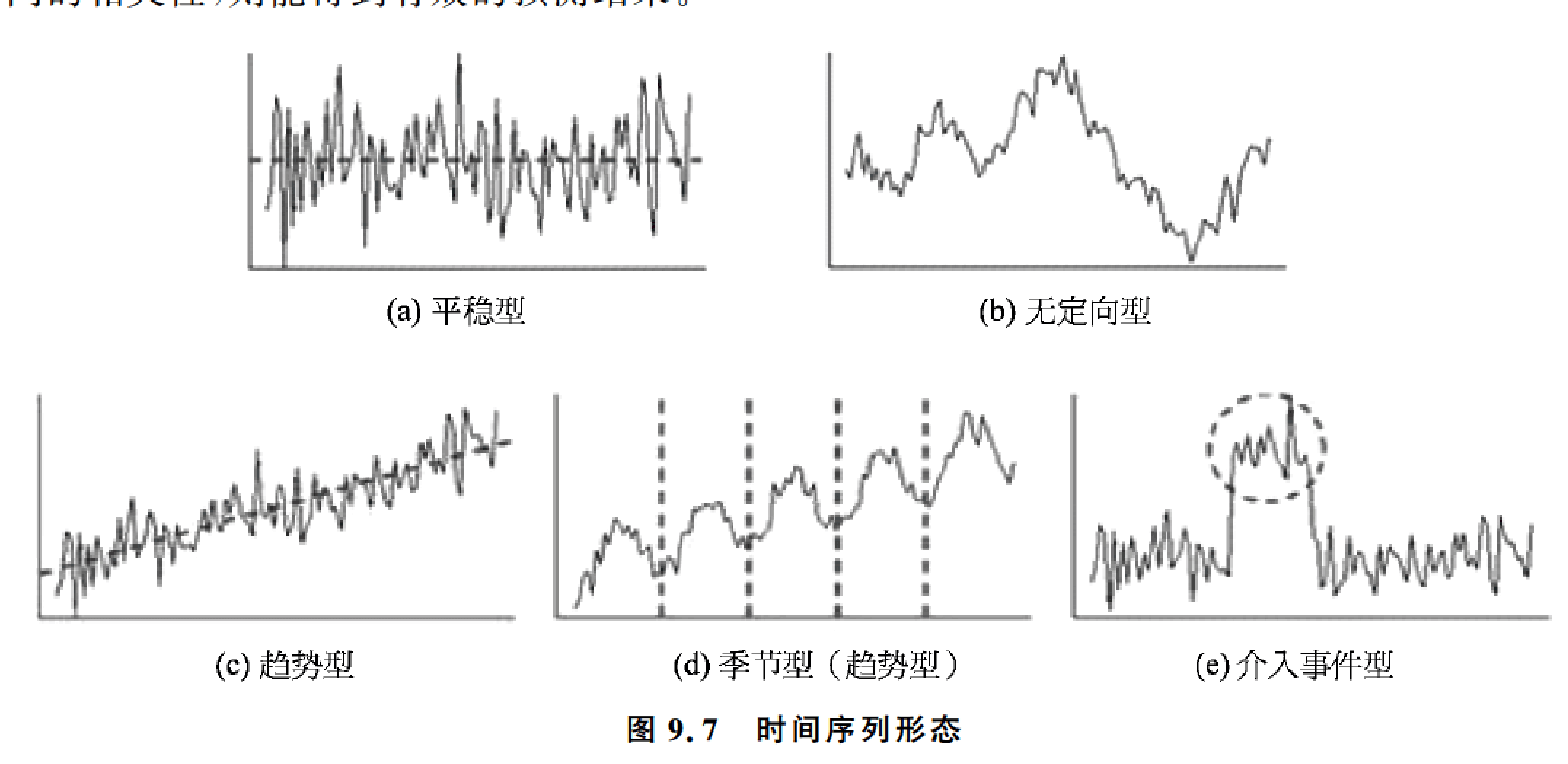


图 9.7 时间序列形态

无定向型时间序列或非平稳型时间序列(non-stationary time series)遇到干扰的时间序列则会呈现波动无定向的状态，如图 9.7(b)所示。外在冲击对序列造成累积的效果，使得序列无法维持固定的水平，因此较难估计预测值。此型序列通常借由差分方程(difference equation)将序列平稳化后再分析。

趋势型时间序列(trend time series)通常是受到长期因素影响，导致序列的平均水平呈现固定趋势变化，但各时间点的数据散布变异固定，如图 9.7(c)。此型序列的平均水平随着时间改变，因此可假设此长期因素将会持续且固定的影响序列，而得出序列预测值。趋势

型时间序列也可借由差分方程将序列平稳化后再行分析。

季节型时间序列(seasonal time series)可以在固定的时间间隔内,观察到类似的波动,如图 9.7(d)为同时具有季节与趋势因素的时间序列。由于此型序列的平均水平有周期性的变动,因此可假设此周期因素将会持续且固定的影响序列,而得出序列预测值。季节型时间序列的预测模式,需要同时考虑观测值之间的相关性与周期性。

介入事件时间序列(interventions time series)因为受到单一的突发事件干扰,而造成序列中少数观测值的表现异于其他观察值,如图 9.7(e)。由于此型序列的平均水平并不变动,且单一突发事件往往无法预测,因此可假设此序列将会维持平均水平与变动,而得序列预测值。离群值时间序列模式须特别加入介入事件参数,以防止单一事件值造成模式的严重偏误。

9.4　时间数据的分析步骤

时间序列分析法可分为时间定义域分析法(analysis in time domain)和频率定义域分析法(analysis in frequency domain)两大类。前者利用**自相关函数(autocorrelation function, ACF)**以建立模式,较着墨于模式构建、参数估计和数据的拟合优度检定,其推导过程仅需适中的观测值;后者以频谱(spectrum)作为分析工具,着重于时间序列的频谱密度及频率范畴分解,其分析结果常被视为系统中基本的变动。以下主要讨论时间定义域分析法,其主要概念是以自相关函数与交叉相关函数(cross correlation function)作为建立随机时间序列模式的依据,并应用所建立的模式进行预测分析。

博克斯和詹金斯(Box & Jenkins,1976)提出时间序列模式构建的试误递归过程(trial and error iterative process),如图 9.8 所示。第一步为了解问题的本质与分析目的,以提升解决问题的效果;第二步为数据准备,包括搜集与检查历史数据、处理遗漏值、转换数据形态、合并或分割数据集合;第三步为观察时间序列的形态,对数据进行检查后依序排

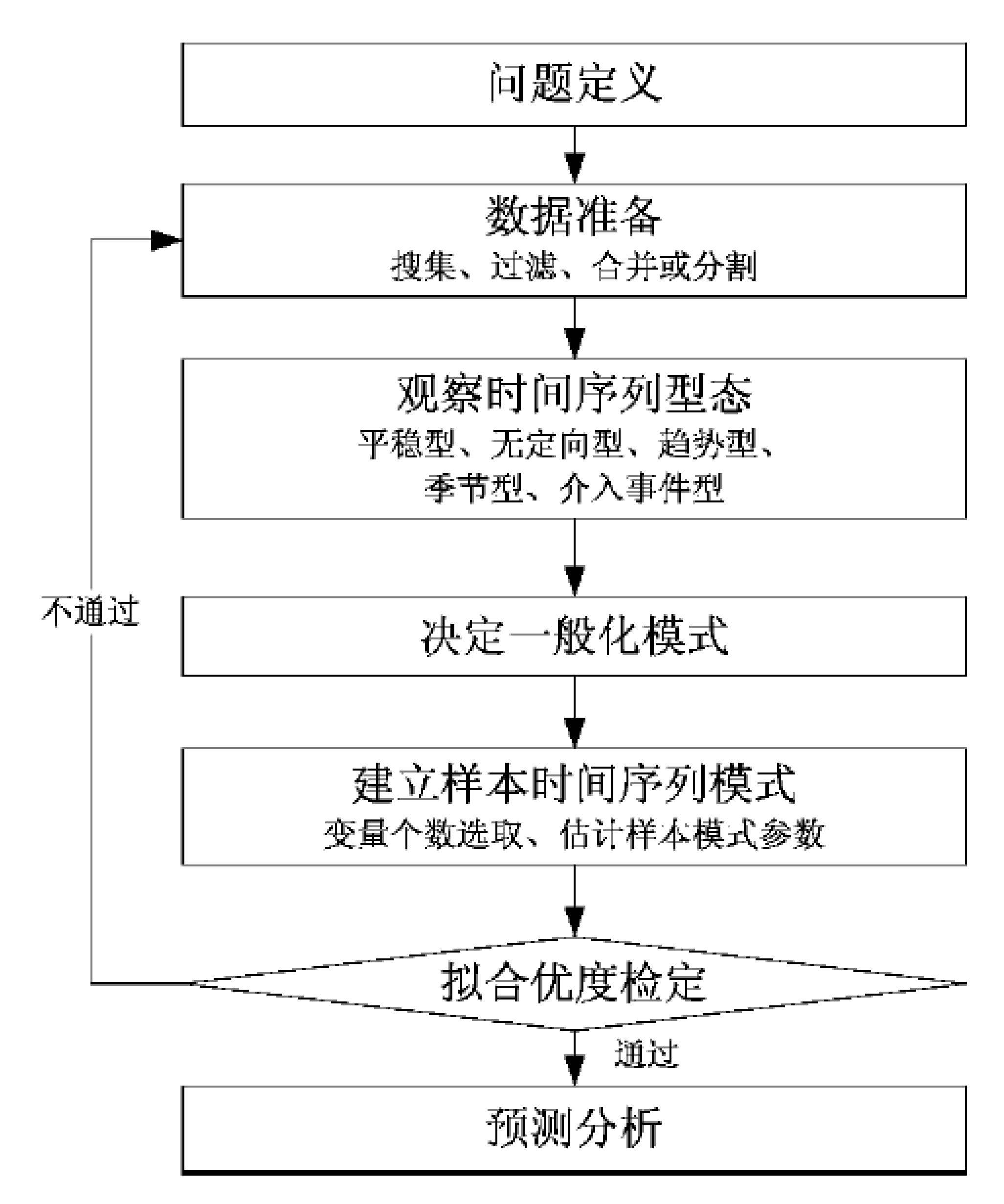


图 9.8　时间序列分析步骤(Box & Jenkins, 1976)

列，计算其自相关函数与**偏自相关函数**(**partial autocorrelation function，PACF**)，并参考图形的特性，提出各种所拟采用的候选模式；第四步寻找时间序列随时间变化的规律，以选取一个合适并精简的模式；第五步利用搜集的数据建立一合适的时间序列模式，其中包含变量个数选取及估计。最后，在进行预测前，必须先诊断所建立的模式与数据的拟合优度检定。若检定结果不通过，则必须重新估计与诊断，直到能获得适当的模式为止。

9.5 模式选择与建立

自相关函数(ACF)与偏自相关函数(PACF)经常搭配使用以检验时间序列形态。自相关函数类似皮尔逊相关系数(Pearson correlation coefficient)，差别在于自相关函数所探讨的为同一变量于不同时期的相关程度，并非不同变量之间的相关性。假设 Z_i，$(i=1,2,\cdots,n)$ 为时间序列的 n 项观测值，相隔 k 期的两观测值的自相关函数可如式(9.14)表示：

$$\rho_k = \frac{\mathrm{Cov}(Z_t, Z_{t+k})}{\sqrt{\mathrm{Var}(Z_t)} \cdot \sqrt{\mathrm{Var}(Z_{t+k})}} \tag{9.14}$$

定义 $\sigma_0 = \mathrm{Var}(Z_t)$ 与 $\sigma_k = \mathrm{Cov}(Z_t, Z_{t+k})$，则 σ_0 与 σ_k 的估计式分别如下：

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

$$\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n-k}(Z_i - \bar{Z})(Z_{i+k} - \bar{Z}) \tag{9.15}$$

其中，$\bar{Z} = \sum_{i=1}^{n} Z_i \Big/ n$ 为$\{Z_i\}$序列的样本平均数。因此，ρ_k 的估计式可构建如下：

$$\hat{\rho}_k = \frac{\hat{\sigma}_k^2}{\hat{\sigma}_0^2} \tag{9.16}$$

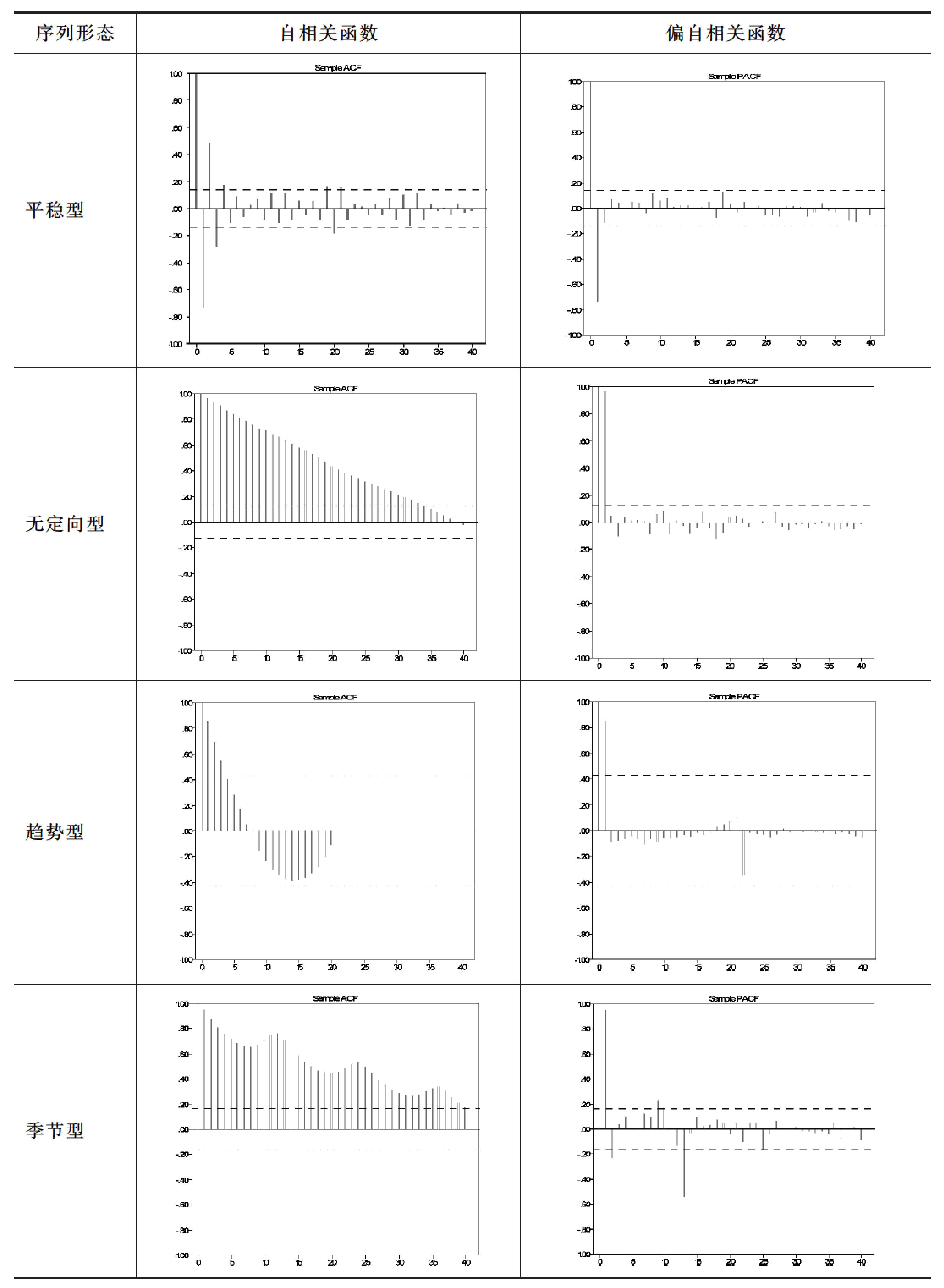
k 期偏自相关函数(PACF)ρ_{kk}是在移除 $Z_{t+1},\cdots,Z_{t+(k-1)}$ 的线性相关下，Z_t 与 Z_{t+k} 两观测值的线性相关程度；与自相关函数的差别在于偏自相关函数是条件相关。利用 1 期与 2 期的自相关系数，2 期偏相关系数的定义如式(9.17)：

$$\rho_{22} = \frac{\hat{\rho}_2 - \hat{\rho}_1^2}{1 - \hat{\rho}_1^2} \tag{9.17}$$

其中，$\hat{\rho}_1$ 与 $\hat{\rho}_2$ 为式(9.16)中，分别将 k 以 1 及 2 代入所得的结果。

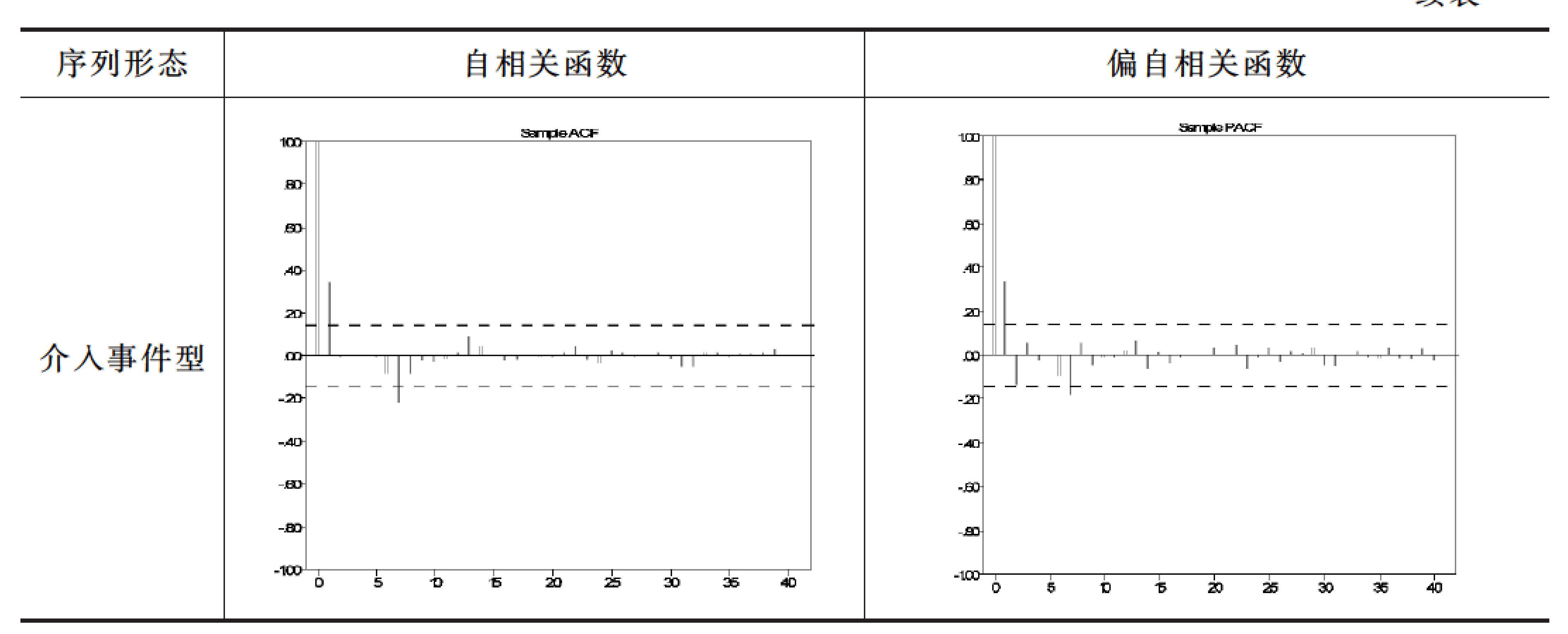
一般而言，平稳型时间序列的自相关函数与偏自相关函数皆会随着时差增加逐渐消失；或在某一特殊时差后，观测值之间的相关性呈现切断的趋势。巴特利特(Bartlett，1937)进而提出用渐进方法来鉴定平稳型时间序列的 ρ_k 是否为 0。表 9.5 说明五种时间序列形态的自相关函数与偏自相关函数的函数图形。

表 9.5　各种序列形态的自相关函数与偏自相关函数的函数图形



序列形态	自相关函数	偏自相关函数
平稳型	Sample ACF	Sample PACF
无定向型	Sample ACF	Sample PACF
趋势型	Sample ACF	Sample PACF
季节型	Sample ACF	Sample PACF

续表

序列形态	自相关函数	偏自相关函数
介入事件型		

9.5.1 时间序列平滑法

移动平均法（moving average method）、加权移动平均法（weighted moving average method）以及指数平滑法（exponential smoothing method）是三种常被用来消除时间序列短期变动的平滑方法，经由选择适当的平滑参数，将数列平滑化后产生的平滑函数，使得序列的长期效应更加明显。前两种的平滑参数为时间间隔长度（time window length），指数平滑法的平滑参数为记忆退化率。以图 9.9 的移动平均法为例，平滑参数越大函数越平滑；反之亦然。

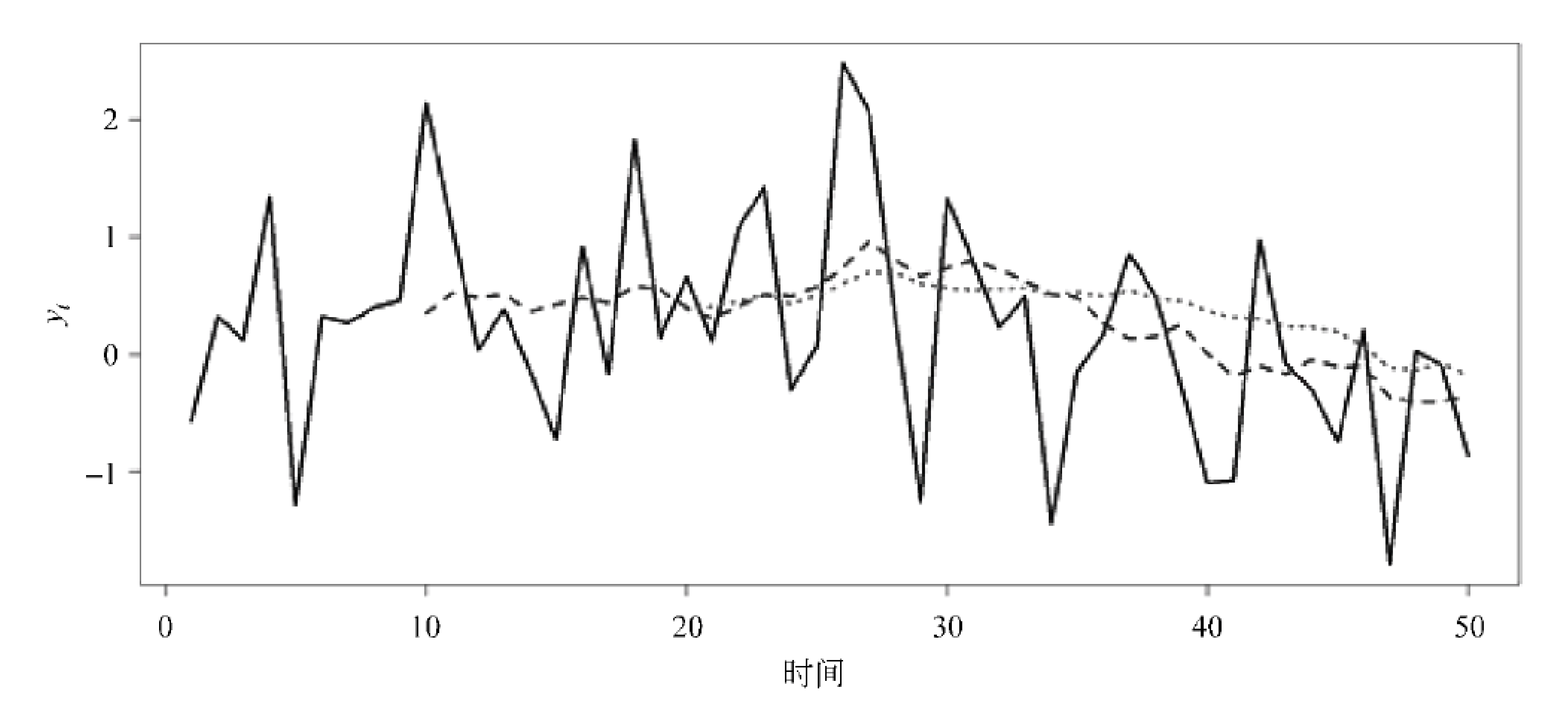


图 9.9　移动平均法在不同时间间隔参数下产生的平滑函数

移动平均法需要决定的只有时间间隔长度参数 k。产生的平滑函数如式(9.18)：

$$y_t = \begin{cases} \dfrac{1}{k}\sum_{i=1}^{k} X_i, & t = k \\ y_{t-1} + \dfrac{1}{k}(X_t - X_{t-k}), & t > k \end{cases} \tag{9.18}$$

最常见的移动平均法如股市的 k 线。加权移动平均法可视为移动平均法的推广，需要决

定的除了时间间隔长度参数 k，还有 k 期权重 $w_i(i=1,2,\cdots,k)$。产生的平滑函数如式(9.19)：

$$y_t = \sum_{i=t-k+1}^{t} w_i X_i, \quad t \geqslant k \tag{9.19}$$

指数平滑法则需决定记忆退化率 α。产生的平滑函数如式(9.20)：

$$y_t = \begin{cases} X_1 & t = 1 \\ \alpha X_t + (1-\alpha) y_{t-1}, & t > 1 \end{cases} \tag{9.20}$$

此三种方法的适用情形及优缺点如表 9.6 所示。

表 9.6　平滑预测法的比较

	移动平均法	加权移动平均法	指数平滑法
适用情况	各观察点的重要性均等	观察点的重要性不同，可赋予权重区分	当数据形态改变，可利用平滑方式取得变动后的权数
优点	计算简单，可消除不规则变动	计算简单，可显示数据的重要程度	储存数据少，数据改变时，权重改变容易
缺点	需储存大量数据	当数据形态改变时，权数变动不易；权数规定会影响预测结果	权数规定会影响预测结果
期数与权数选取	数据敏感度越高，期数选取要越多	数据敏感度越高，近期权数设定越大	规则变动(随机变异)较大时，平滑指数应取较小，避免预测值过度受误差影响

9.5.2　平稳型时间序列

平稳型时间序列的平均水平不因时间变化而改变，但可依其程度区分为严密平稳与衰落平稳两类型。

严密平稳型时间序列在固定时期内的概率分布不因时间起点改变，亦即无论观测时间往前或往后移动，其概率结构均保持不变，如式(9.21)：

$$f_{Z_1,\cdots,Z_k}(z_1,\cdots,z_k) \overset{\text{def}}{=\!=} f_{Z_{t+1},\cdots,Z_{t+k}}(z_1,\cdots,z_k), \quad t \in \mathbb{N} \tag{9.21}$$

衰落平稳序列的概率分布，仅其一阶动差(平均数)与二阶动差(协方差)不随时间起始点移动而改变，故又称为二阶平稳型时间序列。由于多变量正态概率密度函数可以完全由一阶与二阶动差来说明其特性，故具有正态假设的二阶平稳型过程均符合严密平稳型随机过程的特性。

时间序列分析经常假设序列具平稳性(stationary)。然而，实务上，许多时间序列都不符合此假设，因此需要先对序列进行方差平稳转换(variance stabilizing transformation)，再进行差分。若该转换后的序列符合平稳性要求，则以适当模式进行配适，而模式无法解释的残差必须符合**白噪声过程(white noise process)**，亦即序列随机变量彼此独立且同服从于期望值为 0、方差不随时间改变之正态分布。式(9.22)为一平稳型时间序列过程的一般式，又称为线性过滤器(linear filter)。

$$Z_t = \mu + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots = \mu + a_t + \sum_{j=1}^{\infty} \theta_j a_{t-j} \tag{9.22}$$

其中，Z_t 为时点 t 的观察值；a_{t-j} 代表时点 $t-j$ 的干扰项，$j=0,1,2,\cdots$，故 a_{t-j} 必符合白噪声过程，可表示为 $a_{t-j} \sim N(0,\sigma^2)$；$\mu$ 与 θ_j 为固定参数值，μ 表示序列的平均水平；$\theta_j(j=0,1,$

2,…)为移动平均系数。

若移动平均系数$\{\theta_j\}$为有限(finite)或无限且收敛(infinite and convergent),则时间序列$\{Z_t\}$为固定水平μ的平稳型时间序列;反之,移动平均系数若为发散,则$\{Z_t\}$为非平稳序列。此为较概略的平稳型序列观察方式,在序列检定上,常利用后移运算符(backward shift operator)转换时间序列模式,作为判断序列是否平稳的依据。后移运算符常以符号B表示,其为建立在时差j的两观测值或干扰项的恒等式上,故此种表示方式仅适用于随时间变化的序列数据,如式(9.23):

$$a_{t-j}=B^j a_t,\quad Z_{t-j}=B^j Z_t \tag{9.23}$$

因此,利用后移运算符作为辅助函数转换,可将式(9.22)简化如式(9.24):

$$Z_t=\mu+(B^0+\theta_1 B^1+\theta_2 B^2+\cdots)a_t=\mu+\theta(B)a_t \tag{9.24}$$

其中,$\theta(B)$即为以参数θ_j及后移运算符B^j所建立的转换函数。博克斯和詹金斯(Box & Jenkins,1976)推导出若$\theta(B)=0$所解出的根落于单位圆之外,也就是当$B^j>1$或$B^j<-1$时,序列会收敛而满足平稳性的条件。

1. 移动平均过程

假设线性过滤器如式(9.22)仅前q个系数非零,即当$j>q$时,$\theta_j=0$。此过程可视为以白噪声所建立的移动平均过程(moving average process, MA process)模式,如式(9.25)所示:

$$Z_t=\mu+a_t+\theta_1 a_{t-1}+\theta_2 a_{t-2}+\cdots+\theta_q a_{t-q}=\mu+a_t+\sum_{j=1}^{q}\theta_j a_{t-j} \tag{9.25}$$

式(9.25)为$MA(q)$过程,亦称为q阶移动平均过程(moving average process of order q),模式中$a_t,a_{t-1},\cdots,a_{t-q}$代表时点$t,t-1,\cdots,t-q$的白噪声项;$q$为移动平均阶次参数;$(1,\theta_1,\theta_2,\cdots,\theta_q)$为一有限集合的权数,为移动平均过程的模式参数,亦称为震动影响(shock effect)或记忆函数(memory function)。这些假设表示噪声项将持续影响$t,t+1,\cdots,t+q$等$q+1$个时期后才会消失,而其影响程度可以权数数值$(1,\theta_1,\theta_2,\cdots,\theta_q)$来表示,更可利用后移运算符辅助函数的转换如式(9.26):

$$Z_t=\mu+(B^0+\theta_1 B^1+\theta_2 B^2+\cdots+\theta_q B^q)a_t=\mu+\theta_q(B)a_t \tag{9.26}$$

$MA(q)$的自相关函数在时差q以内的移动平均系数不全为零,而自落后q个时期以后全为零,一般称此自相关函数在时间位差q之后截断(cuts off at lag q);而其偏自相关函数会呈现以指数或正弦形态递减至消失,但持续且非切断。由图9.10为$MA(1)$过程的自相关函数示意图,可以发现在时间点1之后切断,即从第2个时间点开始之后的自相关系数皆为零;而其偏自相关函数却呈现正弦形态逐渐趋近于零。移动平均过程的含义为多个干扰项$a_t,a_{t-1},\cdots,a_{t-q}$的移动线性组合,并非真正的移动平均,可由移动平均系数和不等于1得到印证。

2. 自回归过程

回归分析是以一个以上的独立变量预测单一相依变量的表现,着重于探讨独立变量与相依变量之间的关联,倘若将回归分析视为一种预测方法,则可视欲预测的时间序列为相依变量。时间序列的自回归过程中,每个时期的时间序列数据必须同时扮演独立变量与相依变量的角色,也就是将随机过程中任一当期值视为回归模式中的相依变量,而将其前p期



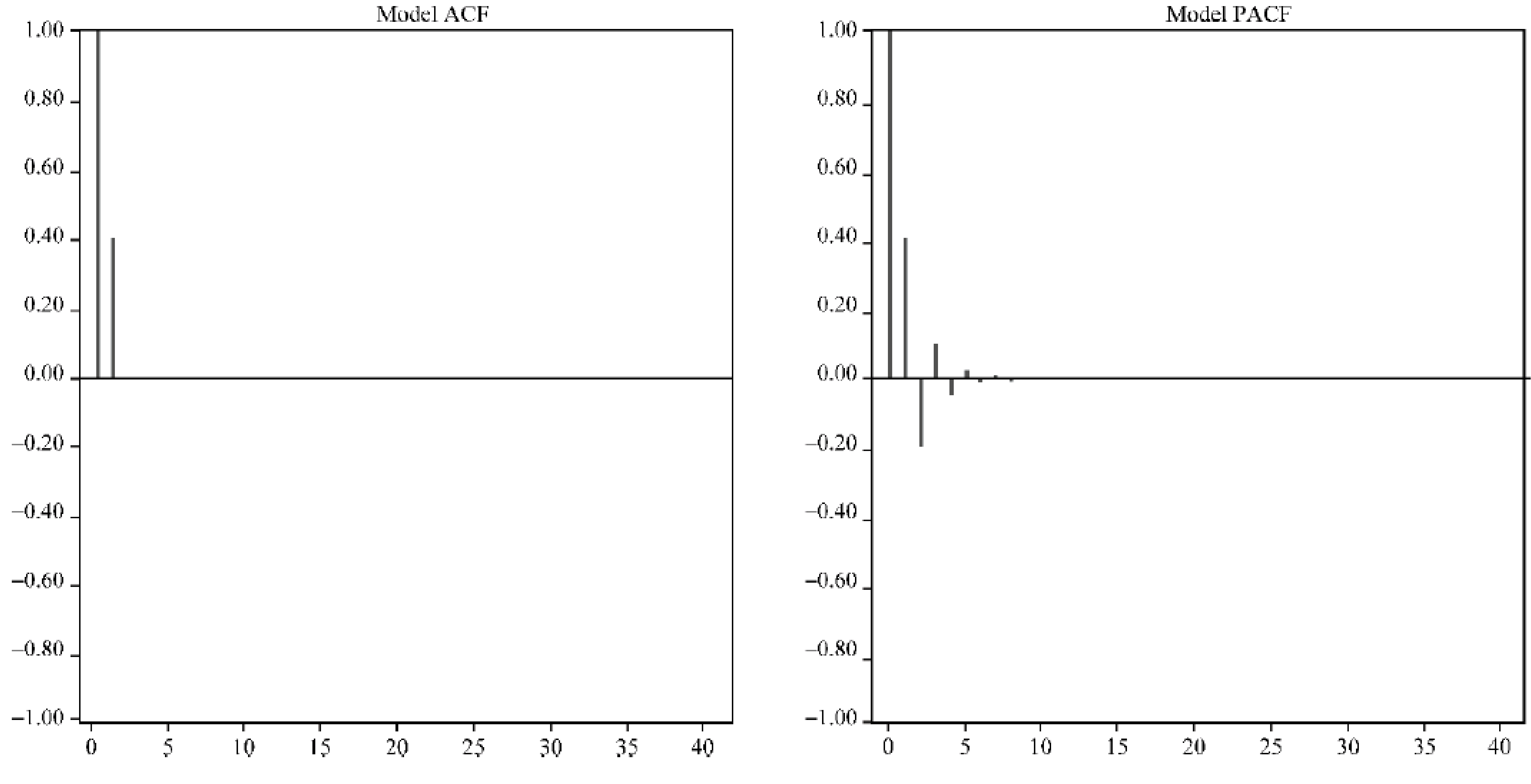


图 9.10　*MA*(1)过程的自相关函数与偏自相关函数

的值视为独立变量来构建回归模式。由于独立变量与相依变量来自同一序列数据,因此称为**自回归过程(autoregressive process, AR process)**,即是以序列的前期值作为独立变量来对预测分析当前值。式(9.22)的观察值 Z_t 受到当期 a_t 与所有过去 a_{t-j} 所干扰,倘若欲了解历史数据对于现在及未来的影响层面,可通过移项与递归的方式,以转化为一种类似回流的线性过程。该转化以当期的干扰与所有过去观察值来表示,为介于回归模式与线性过滤器所发展出的预测模式。

自回归过程常应用于平稳型序列分析上,如式(9.27)即为建立于当期干扰与过去 p 期观测值的自回归模式。

$$Z_t = C + a_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} = \mu + a_t + \sum_{i=1}^{p} \phi_i Z_{t-i} \qquad (9.27)$$

式(9.27)为 $AR(p)$过程,亦称为 p 阶自回归过程,其中,Z_{t-i} 为序列中时间点 $t-i$ 的观察值,$i=0,1,\cdots,p$;参数 p 代表会对现在数值产生影响的过去观测值个数;a_t 为当期的干扰项,符合白噪声过程;ϕ_i 为时间序列模式中待估计参数,是自回归系数,代表过去的数值对现在数值的重要性,亦可利用后移运算符辅助函数的转换为式(9.28):

$$(1 - \phi_1 B^1 - \phi_2 B^2 - \cdots - \phi_p B^p) Z_t = C + a_t \rightarrow \phi_p(B) Z_t = C + a_t \qquad (9.28)$$

$AR(p)$的自相关函数经推导得知会呈现以指数下降趋势,其偏自相关函数当时差小于或等于 p 时不为 0,但大于 p 后皆为 0,即在时差 p 之后截断。图 9.11 显示 $AR(1)$的自相关函数与偏自相关函数的形态,可发现前者以指数形态递减终至为 0;而后者于时间点 1 之后切断,即从第 2 个时间点开始的偏自相关系数为 0,成截断形式。

自回归过程中的基本假设是残差之间彼此独立,且同来自平均为 0 且方差为定值的正态分布,又称为白噪声过程。$AR(p)$过程如同模式(9.28),可被解释为 Z_t 的分解,一部分完全依赖 $\phi_1 Z_{t-1}, \phi_2 Z_{t-2}, \cdots, \phi_p Z_{t-p}$ 而定;另一部分则与 Z_t 无关,由 a_t 决定。当时间点 t 的观察值 Z_t 已知时,a_t 不再是随机变数而是一个定值。

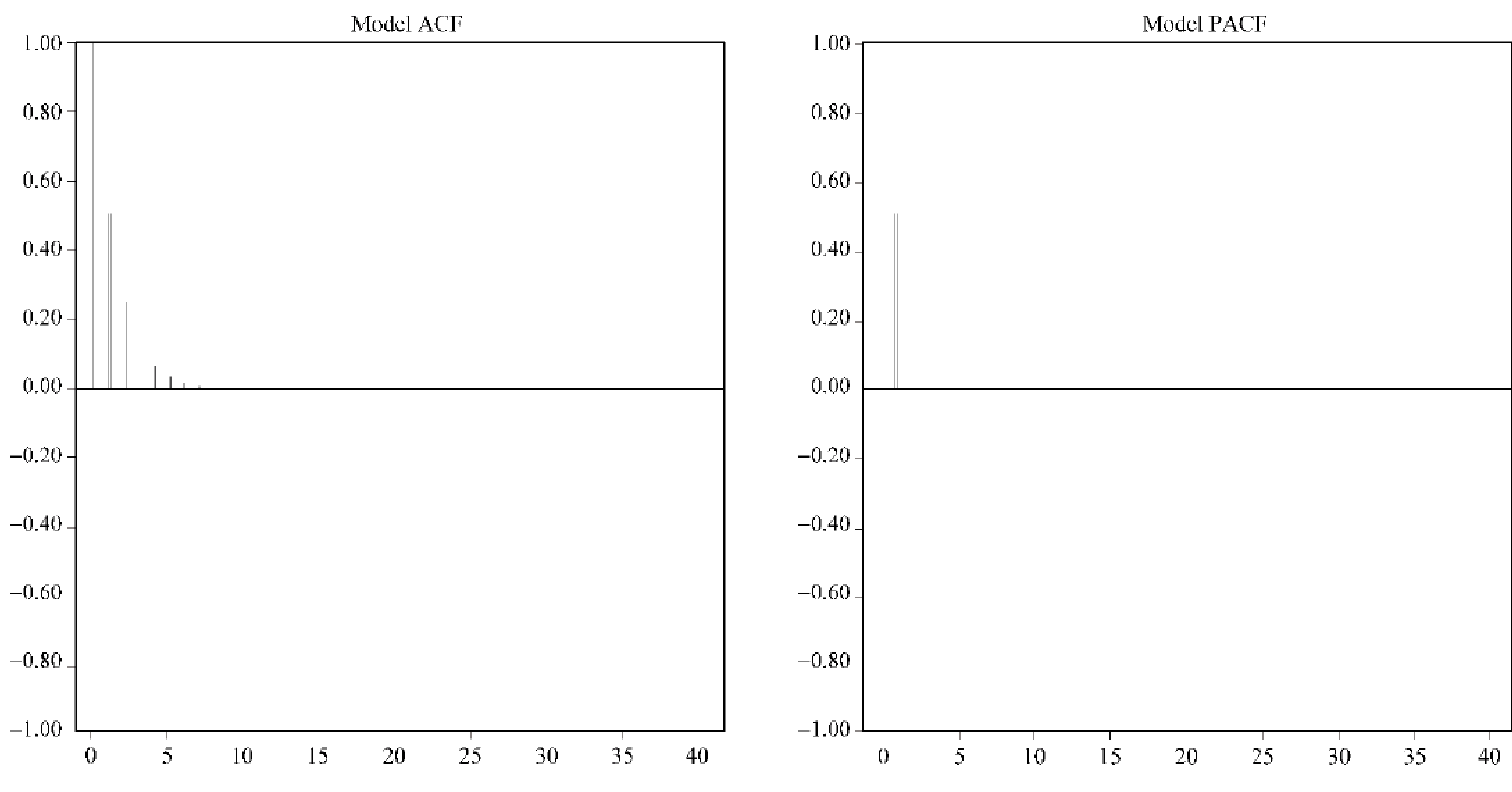


图 9.11 *AR*(1)过程的自相关函数与偏自相关函数

3. 自回归移动平均过程

移动平均与自回归过程具有双重性(duality),在符合限制的移动平均系数与自回归系数下,该过程具有可逆性。如 $AR(1)$过程可转换为 $MA(\infty)$过程,且 $MA(1)$过程亦可转换为 $AR(\infty)$过程。因此,为了精简模式或推导出更贴近实际的模式,可将自回归与移动平均模式结合运用,称为**自回归移动平均过程(autoregressive moving-average process, ARMA process)**。

在 $AR(p)$过程中,可将独立变量 Z_t 分解为两部分,一部分相依于 $\phi_1 Z_{t-1}, \phi_2 Z_{t-2}, \cdots, \phi_p Z_{t-p}$;另一部分为与 Z_t 无关的 a_t 残差项。在 $MA(q)$过程中,Z_t 的预测值全来自于 $a_t, \theta_1 a_{t-1}, \cdots, \theta_q a_{t-q}$所给予的信息。因此当时间序列的数据特性已无法仅用 $AR(p)$过程或 $MA(q)$过程来描述时,可利用合并方式将模式改写为 ARMA 过程,式(9.29)为 $ARMA(p,q)$过程的一般式。

$$Z_t = a_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{9.29}$$

整理式(9.29)可了解自回归过程与移动平均过程对自回归移动平均过程的影响,如式(9.30)所示,等号左边为自回归部分,右边则为移动平均部分。

$$Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} - \cdots - \phi_p Z_{t-p} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{9.30}$$

ARMA 过程可视为自回归过程与移动平均过程的整合模式,其模式与回归模式相似。在回归模式中,时间点 t 时进入模式仅对 y_t 有影响;当系统从时间点 t 进入时间点 $t+1$ 后,此干扰即会消失。回归模式仅存在独立变量与相依变量间的静态关系;ARMA 过程的条件回归模式亦具此静态关系。

总而言之,若时间点 t 所发生的干扰 a_t 会持续对系统发生影响,而这些动态或记忆显示出数据之间的关系,可以 ARMA 模式来描述此系统,并可利用后移运算符简化描述 ARMA 模式,如式(9.31):

$$\phi_p(B)Z_t = C + \theta_q(B)a_t \tag{9.31}$$

9.5.3　无定向型时间序列

差分自回归滑动平均模型(autoregressive integrated-moving average models,ARIMA)为同时考虑固定与不规则两种影响因素。若影响因素为固定因子,则可借由序列中的过去值来推论序列现在与未来的走向,亦即序列符合 ARMA 过程。另一种不规则因子起源于无法解释的变异,可由 ARIMA 分析模式中的差分阶层估计出,也就是差分后的序列符合 ARMA 过程。

面对无定向型时间序列时,常以差分将序列平稳化,即差分后的序列平均水平固定,而差分后的序列为平稳型序列(Granger & Newbold, 1976),此平稳型序列可用 9.5.2 节的方法进行模式构建。若某序列的样本自相关函数呈极缓慢消失,并且序列图不在固定水平内摆动,则显示此序列为无定向型序列,需先进行差分至序列的自相关函数很快消失为止,如图 9.12 所示。

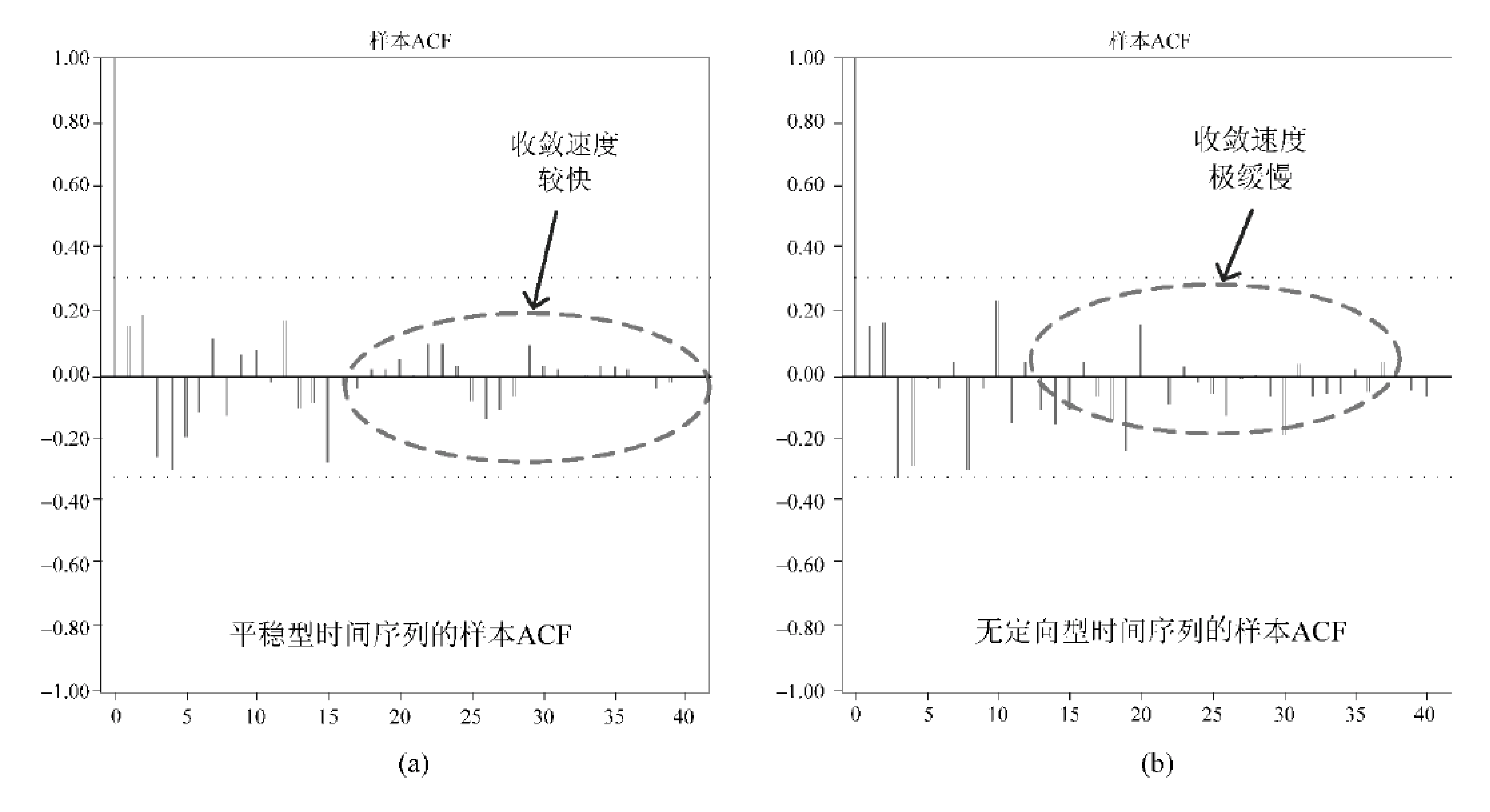


图 9.12　平稳型与无定向型序列之样本自相关函数形态

图 9.13(a)考虑一离散确定型时间序列,属于平均水平与斜率皆随时间递增的无定向型时间序列,可利用一阶差分的动作 $Z_t - Z_{t-1}$ 使其变为仅在水平递增的无定向型序列,如

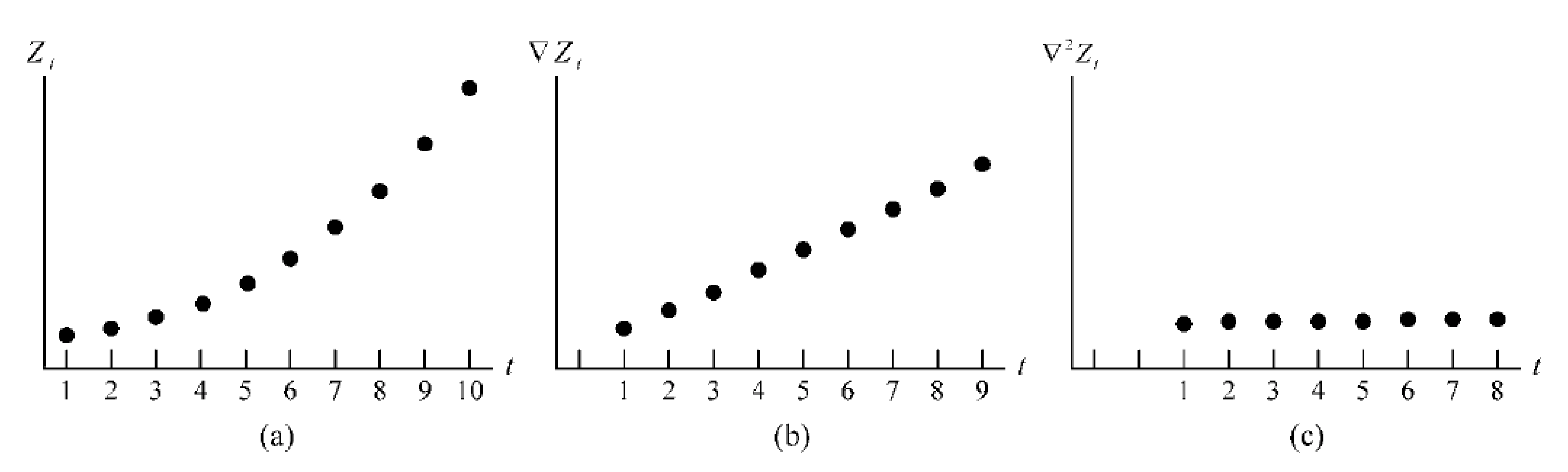


图 9.13　无定向型时间序列的差分转换过程

图 9.13(b)所示。接着,可再取二阶差分$(Z_t-Z_{t-1})-(Z_{t-1}-Z_{t-2})$,使该序列转化为一平稳型时间序列,如图 9.13(c)所示。虽然经由连续差分可以将无定向型序列转为平稳型序列,但差分次数不宜过多,否则将使数据丧失实际含义而不易解释,且使序列的变异变大。实务上常以目测原始序列图形来判断是否已达平稳的状态。

若原始序列经由取 d 阶差分后为 $ARMA(p,q)$过程,则此模式称为(p,d,q)阶整合自回归移动平均模式,记为 $ARIMA(p,d,q)$。转换后的平稳型时间序列不一定为混合型,也可能单纯为 p 阶自回归过程或q 阶移动平均过程,前者称为(p,d)阶整合自回归过程,简称 $ARI(p,d)$或 $ARIMA(p,d,0)$过程;后者称为(d,q)阶整合移动平均过程,简称 $IMA(d,q)$或 $ARIMA(0,d,q)$过程。由 $ARIMA(p,d,q)$过程所产生的时间序列观测值 Z_t,可由多个前期观测值与当期及过去干扰来表示,如式(9.32):

$$Z_t=\phi_1 Z_{t-1}+\phi_2 Z_{t-2}+\cdots+\phi_{p+d}Z_{t-p-d}+a_t-\theta_1 a_{t-1}-\theta_2 a_{t-2}-\cdots-\theta_q a_{t-q} \quad (9.32)$$

式(9.32)与 $ARMA(p,q)$表示方式类似,变量与参数的定义以及残差项 a_t 皆假设符合白噪声。其不同处在于 $ARIMA(p,d,q)$纳入差分项 $\phi_{p+d}Z_{t-p-d}$作为转换无定向型序列至平稳型序列的控件。故当 $d=0$ 时,该过程即为 $ARMA(p,q)$过程。也可将式(9.32)改写,以后移运算符、当期观测值以及当期干扰项表示,如式(9.33):

$$\phi_p(B)\ (1-B)^d Z_t=\theta_q(B)a_t \quad (9.33)$$

ARMA 与 ARIMA 过程最大的特点在于模式仅以过去观测值进行分析与预测,并无独立变量的设定。优点在于不需考虑其他外部数据就可以进行分析,缺点是在数据较复杂的情况下,此模式将不易挑选参数。无定向型序列转为平稳型序列,若起因为平均数为变项,可以 $ARIMA(p,d,q)$过程进行分析;但若起因为方差的变动,则需将原序列经过转换函数(如 Box-Cox 转换等),使其方差为一固定值。一般常见的方法为对原序列取自然对数(natural logarithms)。然而,并非所有无定向型序列都可经由差分或转换函数的方式转为平稳型序列,因此,可以改用自回归条件异方差(autoregressive conditional heteroskedastic, ARCH)模式作为处理会随时间改变的时间序列方差方法(Engle, 1982)。

9.5.4 趋势型、季节型与介入事件型时间序列

时间序列可由加法模型(additive model)与乘法模型(multiplicative model)来表达趋势效应、季节效应以及介入事件效应。加法模型是利用定值的增减以表示趋势及季节所造成的绝对影响,如式(9.34);乘法模型则分别代表趋势与季节的相对影响,通常以平均观测值的百分比表示,如式(9.35):

$$加法模型:Z_t=T_t+S_t+I_t \quad (9.34)$$

$$乘法模型:Z_t=T_t\cdot S_t\cdot I_t \quad (9.35)$$

其中,T_t 为用来捕捉时间序列的趋势效应;S_t 为随着固定时期 s 所变化的函数,用来捕捉时间序列的季节效应;而 I_t 为介入事件效应。因此,若能提取并估计出 T_t,S_t 与 I_t 效应,使其成为已知函数或数值,则更新后的时间序列的残差项即能符合平稳性的随机过程。至此,可利用平稳型概率模式并辅以季节性与趋势效应的信息,构建预测模式。

描述趋势形态的模式约有四种,其函数与图形分别整理于表 9.7。

表 9.7　时间趋势模式

趋势模式	函数式	图形描述	
线性函数	$\hat{Z}_t=\beta_0+\beta_1 t+\varepsilon_t$		
双曲线函数	$\hat{Z}_t=\beta_0+\beta_1 t^{-1}+\varepsilon_t$		
指数函数	$\ln\hat{Z}_t=\beta_0+\beta_1 t+\varepsilon_t$		
二次函数	$\hat{Z}_t=\beta_0+\beta_1 t+\beta_2 t^2+\varepsilon_t$		

加法模型与乘法模型的应用甚广，然而当序列趋势及季节效应非确定值时(例如受到过去观测值间相关性的影响造成有别于白色噪声的波动效果，或是季节效应可能会随着序列的每次循环而有动态性的变化)，则无法仅以 T_t 与定项 S_t 来描述该时间序列的可解释的平方和，改为采用季节性差分自回归滑动平均模型(seasonal autoregressive integrated moving average models, SARIMA)。此模式允许季节效应随着循环而呈现动态变动，非一固定值，式(9.36)为 SARIMA 利用后移运算符的转换后的精简模式。

$$\phi_p(B^s)(1-B^s)^d Z_t = \theta_q(B^s)a_t \tag{9.36}$$

SARIMA 过程的差分模式与 ARIMA 过程式(9.29)极为相似，最大不同在于 SARIMA 过程的模式中，后移运算符的时间差皆为一特定数值 s，此代表每隔 s 个时间间隔的观测值有特定行为或表征产生，故以 B^s 嵌入模式中，以强调序列的季节效应(回顾 $B^s Z_t=Z_{t-s}$ 且 $B^s a_t=a_{t-s}$)。

一般而言，季节事件发生通常有一定的规则和周期性，且常伴随着趋势发生，如图 9.14。若能配适与解释季节变动的规则性，即能善用该信息而使预测模式更加准确。除了可应用 SARIMA 过程的差分模式来建立具有季节效应的时间序列外，一种常用于同时处理季节效应与趋势效应的方法即为上述提及的 $ARIMA(p,d,q)$模式，其做法为针对原时间序列重复应用差分运算直到该转换后的序列呈现平稳型序列分布，再利用 $ARMA(p,q)$的建立方式取得预测模式。

如前所述，差分的运算不宜过多，否则会导致模式解释力不足及失真。趋势及季节的配适同时考虑方差的改变，因此可使模式的解释能力更佳；若强制配适差分模式，虽可消除原

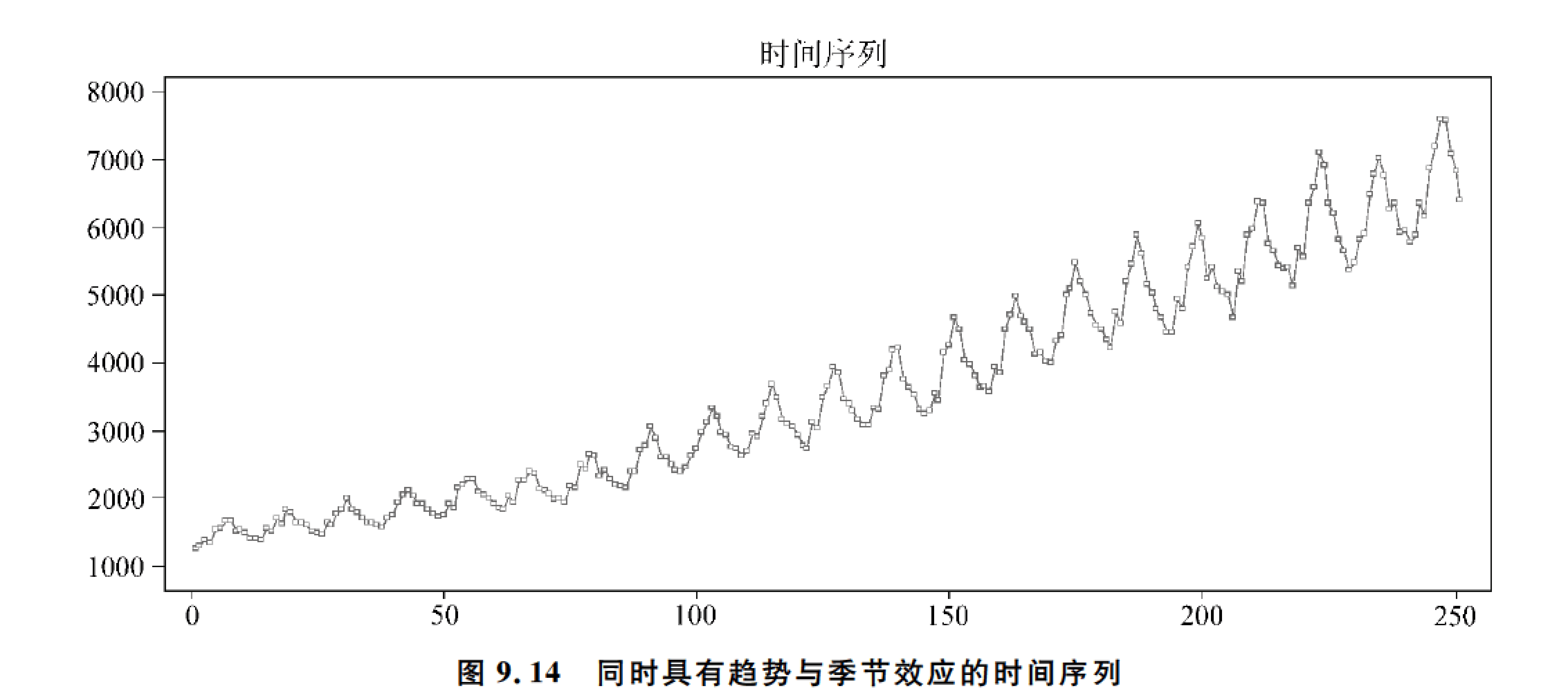


图 9.14 同时具有趋势与季节效应的时间序列

序列趋势与季节效应，但模式预测准确率容易偏低，增加解释模式的困难。因此，若可善用阶段式分解法，先提取季节效应与趋势效应，再针对剩余的序列波动进行模式求解及验证，将可得到更多可用于预测的信息。

时间序列常受到政治、经济、天灾等介入事件影响，造成时间序列的漂移，统称为离群值。假若介入事件的发生时间点可预知，则可应用介入事件模式来捕捉其影响，如以转换函数模式的形式来解释各种假设之间的动态关系。例如，Chien 和 Lin(2012)应用灰预测来估计新竹科学园区的半导体总体产值，以提供个别公司根据其市占率修正其预测，以及作为产业上下游公司之间的领先信号，即另外考虑重大事件的影响。

9.6 阶次选取与参数估计

在处理实际问题时，分析者应了解时间序列的基本结构再选择候选模式，才能有效描述、解释甚至预测时间序列数据。常见的 ARIMA 模式选择，包含变量个数选取、阶次选取(order selection)(即决定 p, d, q 数值)及参数估计(parameter estimation)。变量个数的选取，如同回归分析等统计模式变量选取的概念，随着所选取的变量个数增加，残差平方和会跟着降低(Brockwell & Davis, 1991)。然而，这并不表示分析者能以选择大量的变量来降低构建模式的残差，以免模式过度配适。假若，欲以 $AR(p)$模式配适 100 个观察值所构成的时间序列，选择 $p=99$ 所构建的模式仅能给予此时段良好的预测值；当欲预测未来的观察值时，常会产生相当大的误差，亦即过度配适的模式无法应用于预测上。

用来协助模式变量个数选取的方法包括图形判断和数值检验两种。表 9.5 中，各种时间序列的自相关函数图形与偏自相关函数图形都有其特定的形态，如渐趋消失、于某时间点后截断、长期效应等，都可用于协助选择子模式及变量个数选取。图 9.15 为具有 40 个观察值的时间序列的自相关函数与偏自相关函数图形。由图中可看出其数值均收敛于满足平稳变异的界限内，且函数图形皆为渐进消失而非有限切断型，故极有可能符合平稳型 $ARMA(p,q)$过程。由于 p 的选定取决于偏自相关函数中的显著时间差，也就是在图 9.15 的偏自相关函数图形中，直至时隔三个时期的偏自相关系数显著，而后间隔四个

时期以上的观察值间的相关性均不显著；同样地，q 的选定取决于自相关函数的自相关系数，因此可以 $p=3,q=1$ 的变量个数组合作为子模式建立的基础。

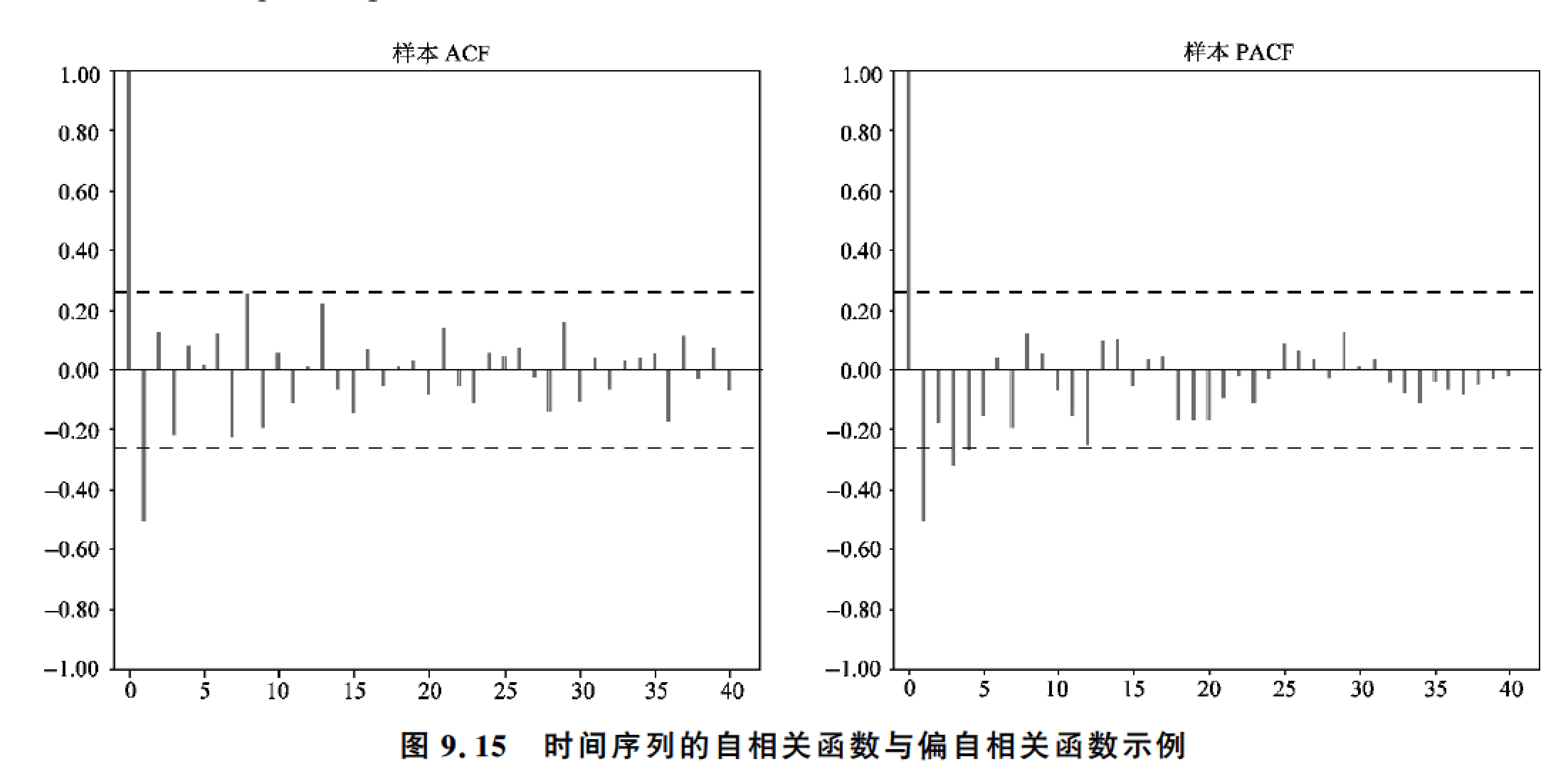


图 9.15　时间序列的自相关函数与偏自相关函数示例

此外，可以用数值检验的方式来决定选取的变量个数，常用的准则有 FPE（finite prediction error）、AIC（Akaike information criterion）以及 BIC（Bayesian information criterion）等，其定义式如表 9.8。

表 9.8　三种检验准则的定义式

准 则	定　义　式	发 展 学 者
FPE	$FPE=\hat{\sigma}^2(n+p)/(n-p)$	Akaike（1969）
AIC	$AIC=-2\ln L(\theta)+2(p+q+1)$	Akaike（1974）
BIC	$BIC=-2\ln L(\theta)+\ln(n)\cdot(p+q+1)$	Schwarz（1978）

表 9.8 中，n 代表观测值个数，$L(\theta)$为该时间序列子模式的最大似然函数。而选取的准则是取递归实验中，使得选定准则最小化的 p 与 q，即为最佳变量个数组合。这三种方式中，FPE 准则仅适用于 $AR(p)$模式中阶次的选取；AIC 准则是以最小化候选模式与实模式间的库克距离（Cook's distance）为目标式而进行 $ARMA(p,q)$阶次选取；BIC 准则是以贝叶斯条件概率寻求最小化库克距离的阶次组合。由各定义式可发现并未提及差分参数 d 个数的选取，原因是并非所有时间序列的模式构建均需使用差分动作，仅于无定向型序列、季节性序列或趋势性序列会采用差分运算后，以求转换一平稳型时间序列，再加以推论，所以一般会先将序列差分为平稳型数列，再用表 9.8 的准则选模。

9.7　模式评估

9.7.1　拟合优度检定

在进行预测前，须先诊断与检定所建立的模式是否适当，例如误差项是否独立或同分布

等。若检定结果显示该建立模式配适得当，则可应用于预测推论；反之，则必须重新找寻适当候选模式、参数估计与模式诊断，直到获得适当模式为止。

当所欲配适模式为 $ARMA(p,q)$ 模式时，需找到适当的变量个数组合并估计母体参数 ϕ、θ 以及 σ^2，并对于每个时期 t 求得其预测值 $\hat{Z}_t(\hat{\phi},\hat{\theta})$，以计算其预测值与实际观察值之间的误差，其残差项 W_t 定义如式(9.37)(Ansley，1979)：

$$W_t = \frac{Z_t - \hat{Z}_t(\hat{\phi},\hat{\theta})}{\sqrt{r_{t-1}(\hat{\phi},\hat{\theta})}}, \quad t = 1,2,\cdots,n \tag{9.37}$$

其中，$r_{t-1}(\hat{\phi},\hat{\theta})$ 为第 t 时期的实际值与预测值的协方差，如式(9.38)：

$$r_{t-1}(\hat{\phi},\hat{\theta}) = E[Z_t \cdot \hat{Z}_t(\hat{\phi},\hat{\theta})]/\sigma^2, \quad t = 1,2,\cdots,n \tag{9.38}$$

若所构建的 $ARMA(p,q)$ 模式适当，则残差应为一白噪声过程，亦即 $\hat{W}_t \sim WN(0,\hat{\sigma}^2)$，如图 9.16 所示，其残差值散布情形如一平稳型时间序列。然而，残差项彼此间的相关结构很难从图形辨别；因此，欲检验所建立模式是否恰当，还需经过假设检定来辅助判断残差项是否服从白噪声过程。

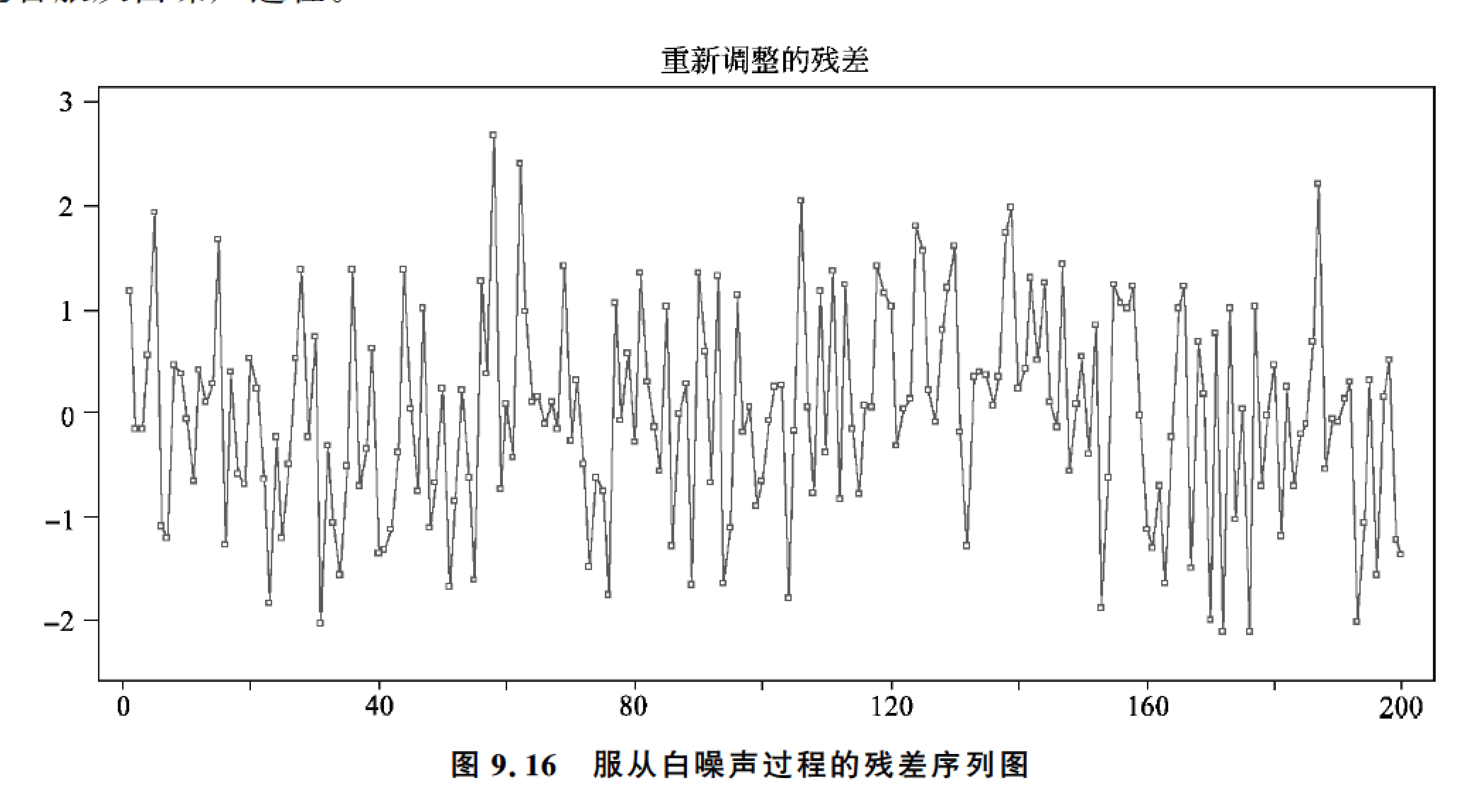


图 9.16 服从白噪声过程的残差序列图

博克斯和皮尔斯(Box & Pierce，1970)提出以近似卡方分布的 Q 统计量，如式(9.39)，以检定“残差项服从白噪声过程”的虚无假设，计算不同时间间隔自相关系数而得一自相关函数矩阵，以求得 Q 统计量，并与卡方临界值比较，以推论是否拒绝虚无假设。

$$Q = n\sum_{j=1}^{h}\hat{\rho}_j^2, \quad h = \{p+q; h \geqslant j\} \tag{9.39}$$

其中，n 为观察样本个数；$\hat{\rho}_j$ 为 j 个时间间隔之估计残差自相关系数；h 为 ARMA 过程中的变量个数；在虚无假设成立下，Q 统计量会近似于 $\chi^2(h)$ 分布。

9.7.2 预测误差衡量

为了比较不同时间序列模式间的准确性，并选择预测能力较佳的模式，在模式建立阶段

一般会尽可能地将模式误差最小化，并借由模式解释能力的高低（可以 R^2 反应模式解释能力）、参数估计值的显著性以及模式预测能力强度（可由 RMSE、MAPE 等指标评估其预测能力），选取最适的趋势模式。

假设有共有 k 期的观察值 $y_t(t=1,2,\cdots,k)$ 与其预测值 f_t，则模式的预测误差如式(9.40)：

$$e_t = y_t - f_t, \quad t = 1,2,\cdots,k \tag{9.40}$$

常见的时间序列数据模式的比较可应用的衡量指标包括平均绝对误差（mean absolute error，MAE）、均方误差（mean squared error，MSE）、平均绝对百分误差（mean absolute percentage error，MAPE），如式(9.41)至式(9.43)：

$$\text{MAE} = \frac{1}{k}\sum_{t=1}^{k}|y_t - f_t| \tag{9.41}$$

$$\text{MSE} = \sum_{t=1}^{k}\frac{(y_t - f_t)^2}{k} \tag{9.42}$$

$$\text{MAPE} = \frac{1}{k}\sum_{t=1}^{k}\frac{|y_t - f_t|}{y_t} \cdot 100\% \tag{9.43}$$

MAE 为所有误差的绝对值平均，主要用以衡量预测的误差大小，假设每笔误差的影响均相同，并不考虑其高估或低估，MSE 则计算所有数据的残差平方和，有时会以取 MSE 的根号为均方根误差（root mean squared error，RMSE）。与 MAE 不同的是，MSE 会将误差放大，使得差异更明显。另外，如果数据间的误差值变化范围很大时，则使用 MAE 或 MSE 可能会造成误判，例如一组数据的实际值与预测值为（10，9），另一组数据为（10 000，9999），虽然误差值都是 1，但第一组数据的误差 1 占原实际值的比例为 10%，而第二组数据的误差则为 0.01%。欲避免此情况，可改用 MAPE 作为误差衡量指标。

9.8　R 语言与时间数据分析

本节以内建在 R 语言中的禾本科植物吸收二氧化碳时间序列数据集（Carbon Dioxide Uptake in Grass Plants）为例，包括自 1959 年 1 月至 1997 年 12 月间的月数据，共计 468 笔观测值（Pinheiro & Bates，2000；Potvin et al.，1990）。为能有效进行分析引入两个实用的扩充套件，分别为 **TSA**（Cryer & Chan，2008）与 **forecast**（Hyndman & Khandakar，2008）。

首先，将数据集依照时间顺序画出趋势图以初步判断是否为平稳序列，亦可借由自相关函数图来辅助判断。扩充套件 **forecast** 中的 **tsdisplay** 函数可将一组时间序列数据同时画出趋势图、自相关函数（ACF）图与偏自相关函数（PACF）图，如图 9.17 所示。一个平稳序列的条件为平均数与方差为与时间无关的固定常数，而由图 9.17 上方的趋势图可明显看出平均值随着时间而变化，变异则较无明显的随时间改变，同时左下方的自相关函数图并无随着时间位差（lag）变大而有截断（cut off）或明显的递减趋势。因此，若要建立此时间序列模式，需先经过差分转换为平稳序列。

```
library(forecast)
library(TSA)
data(co2,package="datasets")
```

```
tsdisplay (co2)
```

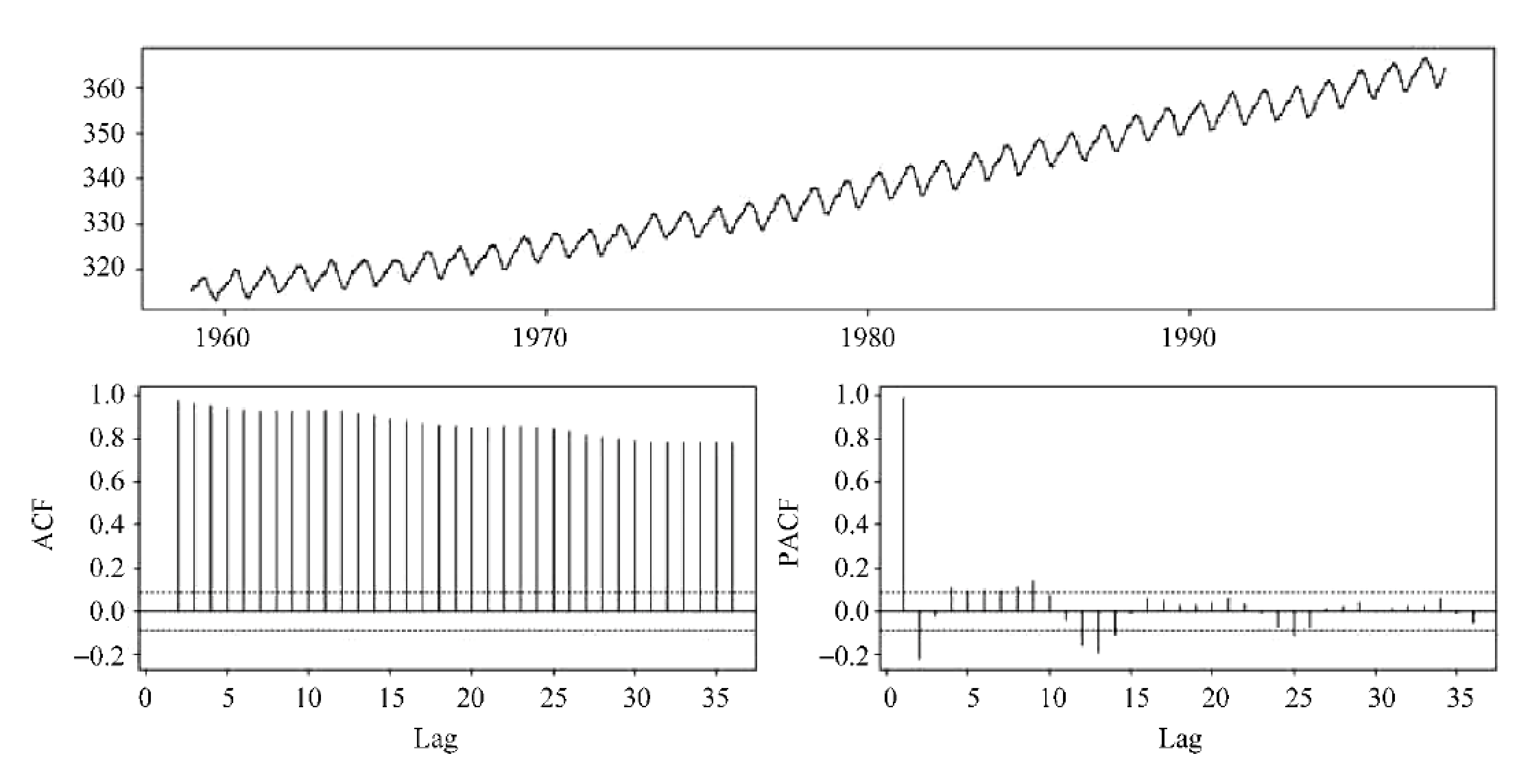


图 9.17　二氧化碳数据图形检查

接着,将此数据集作分割,保留最后 12 笔月数据作为验证数据,并将训练数据进行一阶差分后再次检查图形。由图 9.18 上方的趋势图可看出数据经过一阶差分后平均值已呈现平稳状态,自相关函数图的时间位差变化也有所改善,但其季节性时间位差(lag=12, 24, 36)的相关性仍十分明显,需进一步作季节性差分才能转换成平稳序列,如图 9.19 所示。

```
train= ts(co2[seq(1,length(co2)- 12)], frequency= 12,start= c(1959,1))
test= ts(co2[seq(length(co2)- 11,length(co2))], frequency= 12,start= c(1997,1))
tsdisplay(diff(train), main= "First difference of co2")
tsdisplay(diff(diff(train),lag= 12), main= "First and seasonal difference of co2")
```

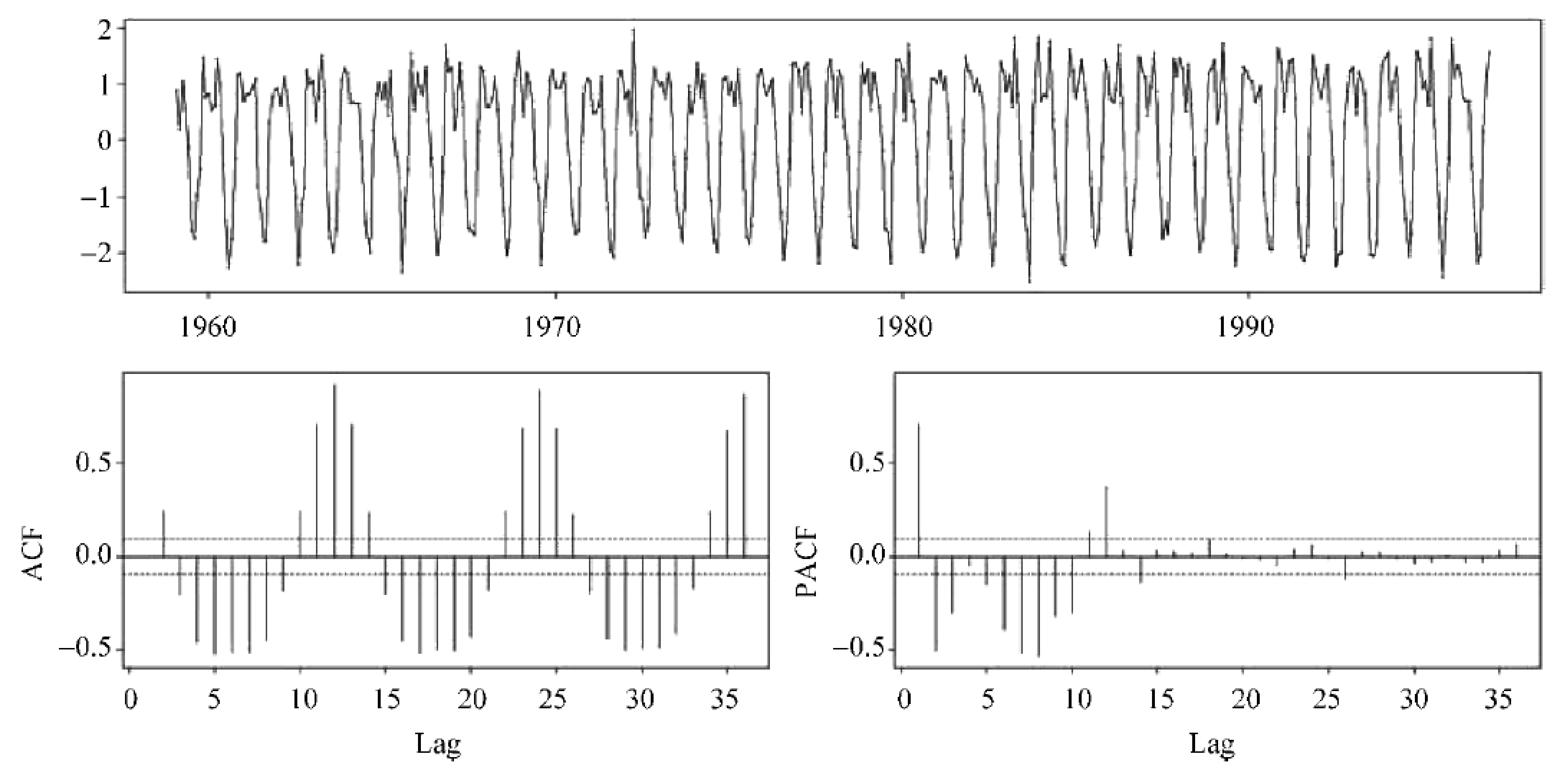


图 9.18　二氧化碳数据一阶差分图形检查

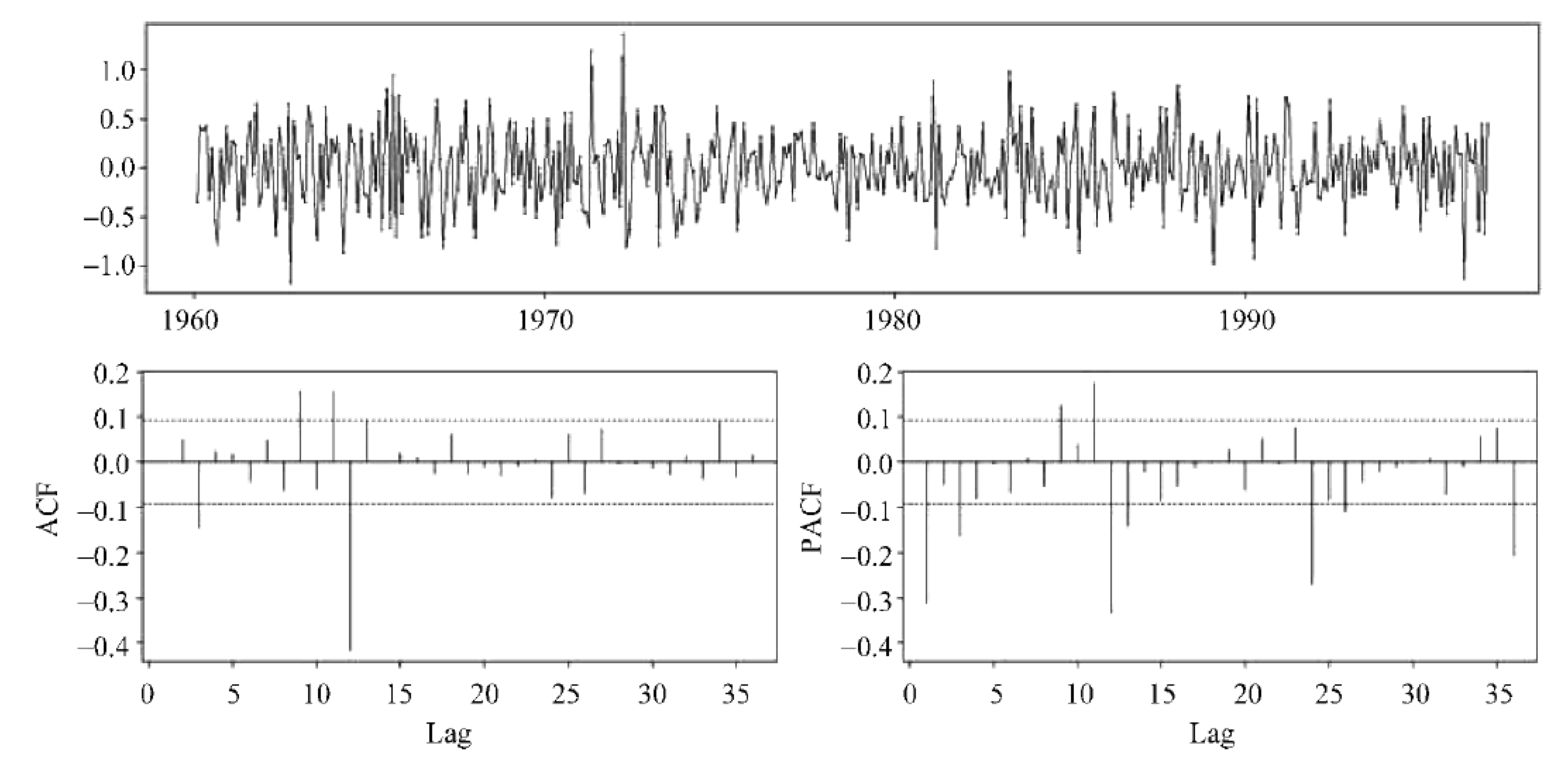


图 9.19　二氧化碳数据一阶差分与季节性差分图形检查

将数据转换成平稳序列之后，再通过图 9.19 下方自相关函数图与偏自相关函数图的行为变化以初步辨识可能的候选模式。由于此数据具季节性，故需分别判断季节性与非季节性的可能候选模式，再合并进行参数估计。在非季节性部分，偏自相关函数图随着时间位差呈现递减趋势，而自相关函数图则是明显截断于 lag＝3，因此可初步判断非季节型部分为一个 $IMA(1,3)$的时间序列模式；在季节性部分，通过观察季节性时间位差(lag＝12，24，36)的变化，可看出偏自相关函数图形随着时间位差呈现递减趋势，而自相关函数图则是明显截断于 lag＝12，因此可初步判断非季节型部分为一个 $IMA(1,1)_{12}$的时间序列模式；两者合并后成为 $ARIMA(0,1,3)(0,1,1)_{12}$模式，共有 4 个参数需进行估计。

通过以下程序可对模式进行参数估计与模式诊断。表 9.9 与表 9.10 分别为参数估计结果与协方差矩阵，显示 4 个参数中仅有 ma2 参数不显著(估计值的绝对值小于等于两倍标准误)，且参数估计之间的协方差都非常小，代表参数估计值之间不会相互混淆，具备统计上的可信度。

```
m1=Arima(train, order=c(0,1,3), seasonal=list(order=c(0,1,1),period=12))
m1$var.coef
tsdiag(m1,gof=36)
qqnorm(residuals(m1)); qqline(residuals(m1))
legend("topleft",legend=paste("p-value =",
+ round(shapiro.test(residuals(m1))$p.val,4)))
```

表 9.9　二氧化碳数据模式参数估计

	ma1	ma2	ma3	sma1
估计值	−0.3356	−0.0096	−0.1102	−0.8572
标准误	0.0480	0.0504	0.0470	0.0260

表 9.10　二氧化碳数据模式参数估计协方差矩阵

	ma1	ma2	ma3	sma1
ma1	0.0023	−0.0008	−0.0001	−0.0002
ma2	−0.0008	0.0025	−0.0008	0.0000
ma3	−0.0001	−0.0008	0.0022	0.0001
sma1	−0.0002	0.0000	0.0001	0.0007

图 9.20 为该模式的残差诊断，左方三个图由上至下分别为模式残差序列图、模式残差自相关函数图、模式 Box-Pierce 统计检定结果(以 p-value 呈现)，右方则为模式残差的正态分配 Q-Q 图。由于模式残差自相关函数图中大部分的值均在红线范围内(除了 lag=9 与 lag=34 之外)、Box-Pierce 检定的 p-value 均大于 0.05 且正态性检定 p-value 也大于 0.05，代表此模式的残差服从白噪声过程，为可接受的模式。

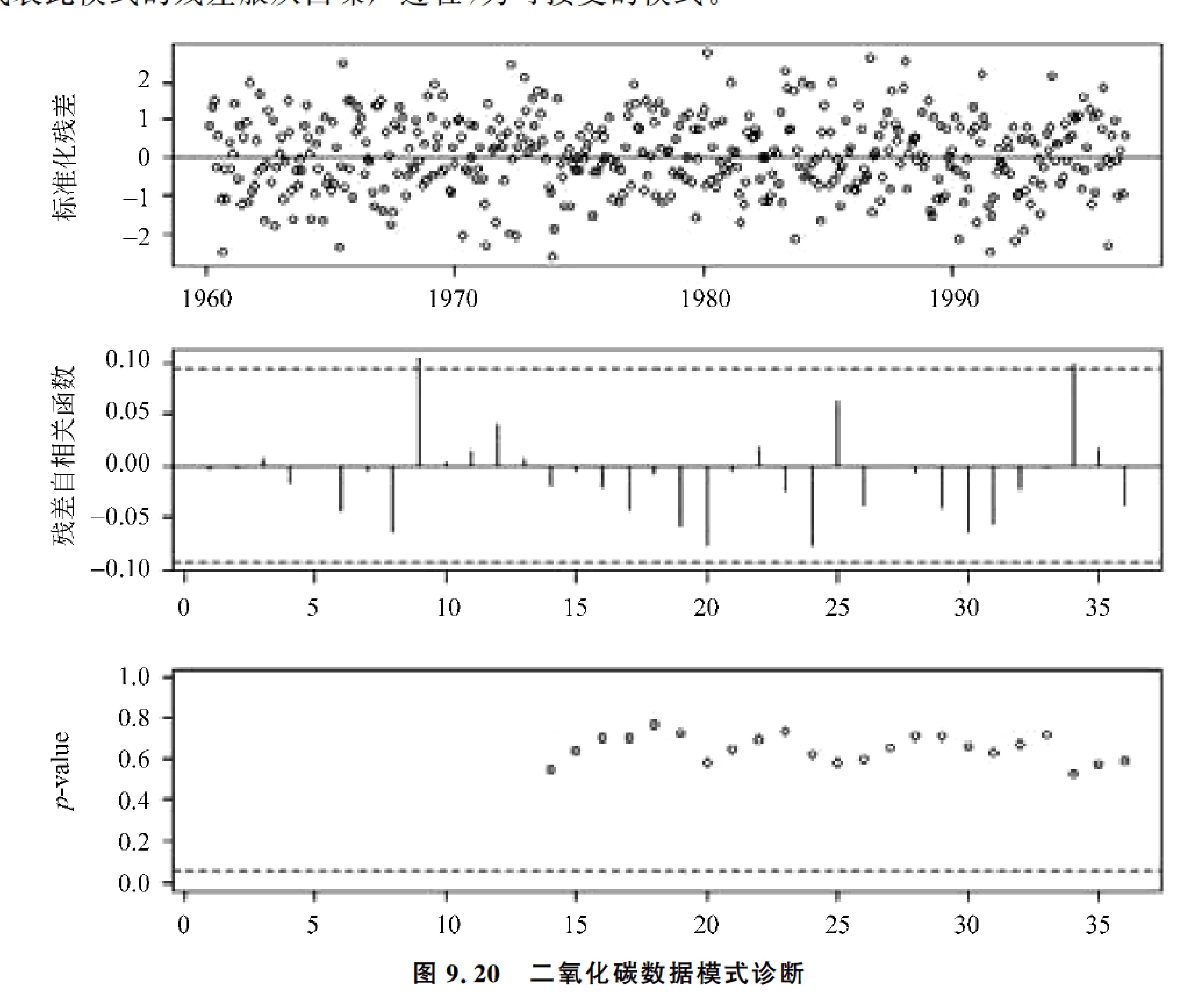


图 9.20　二氧化碳数据模式诊断

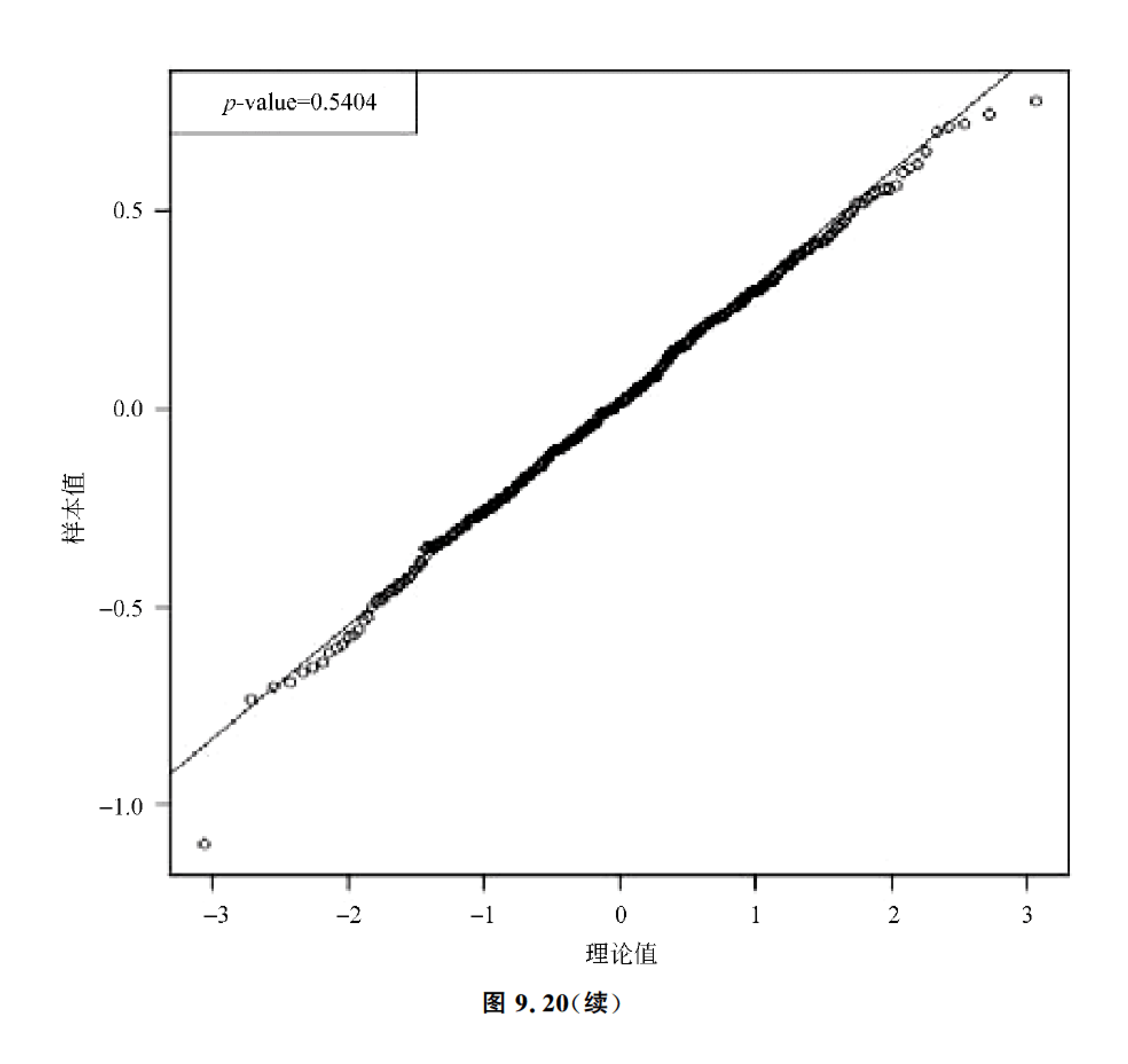


图 9.20(续)

9.9　应用实例——半导体光罩需求预测

9.9.1　案例简介与问题架构

光罩为半导体生产过程重要的零部件,因此掌握光罩的需求可以估计未来该半导体产品的订单需求、协助制定产能分配策略、提升整体获利。然而,半导体制程科技因世代间的差异和市场的变化,所以无法完全使用旧制程的历史数据和需求样型来预测先进制程的未来需求。

本案例(Chien,*et al*.,2010)针对半导体制程光罩订单需求的时间序列数据,发展两阶段半导体制程光罩需求预测模式。第一阶段是用概率密度函数配适,求得该制程光罩需求的趋势,也就是该制程的生命周期;第二阶段则是用过滤序列制程生命周期的数据为输入数据,进行剩余需求波动变化检测,以了解是否隐含其他有价值的信息,并建立时间预测模式以洞悉未来变化。以下说明如何应用时间序列方法分析去除生命周期趋势后的需求波动。

9.9.2　数据准备与数据处理

制程 A 需求波动曲线如图 9.21 所示。在取得制程 A 的生命周期下,将其先由历史数据中移除,并以剩余波动序列为输入数据,如图 9.21 的余波序列即为实际订单数量与生命

周期函数的差分值。本案例以时间序列分析方法作为构建描绘干扰波动的预测模式，其步骤分别为 ARMA 的模式构建与白噪声过程的验证。

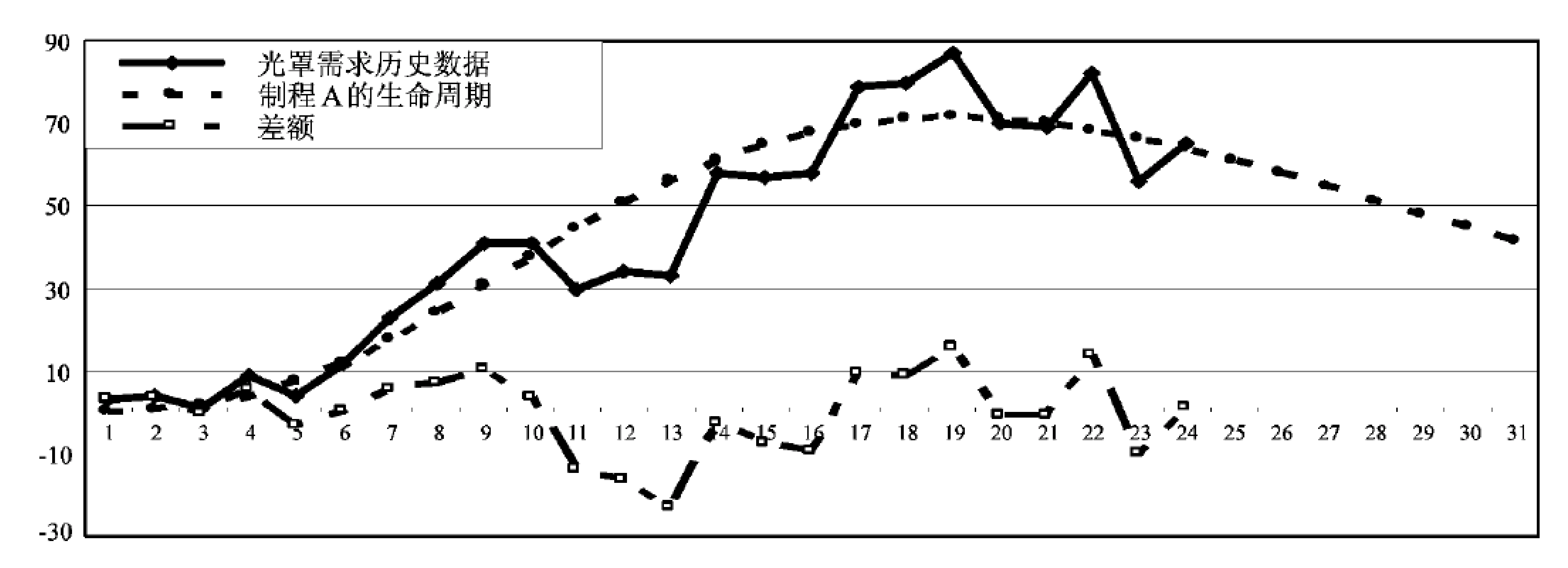


图 9.21　过滤生命周期后的余波序列

9.9.3　需求波动侦测分析过程

本案例采用时间序列分析方法，将制程生命周期趋势移除后，进而分析剩余波动序列是否存在显著的干扰或波动，其中包含自回归移动平均过程的模式构建与白噪声过程的验证两个步骤。

1. 自回归移动平均过程的模式构建

由于顾客订单需求量的各观察值间存在高度相关性，因此时间序列分析可作为良好的模式构建工具，借由本身的历史数据以建立预测未来趋势的模式。而自回归移动平均过程合并自回归过程与移动平均过程以形成动态预测模式，提供更为精确的需求预测以协助后续决策的评估(Box & Jenkins，1976)。

构建 ARMA 模式中主要包括阶次选取与参数估计。分析者可由自相关函数与偏自相关函数得到初步判断的线索，并借由法拉维和查特菲尔德(Faraway & Chatfield，1998)提出的 BIC 准则作为阶次选取工具，以具有最小 BIC 值的(p,q)阶次组合构建 ARMA 模式。决定最佳阶次组合后，再以最大似然估计法进行 ARMA 模式的参数估计。

在 $ARMA(p,q)$模式构建中，首先需决定阶次 p 与 q 的大小。在此，以 BIC 门槛值进行阶次选取，其结果为 $ARMA(6,0)$(或可简记为 $AR(6)$)。接着以最大似然估计法进行参数估计，以构建自回归移动平均过程的模式，其结果如式(9.44)所示。

$$X_t = 0.3264X_{t-1} + 0.01391X_{t-2} + 0.01549X_{t-3} - 0.2504X_{t-4} + 0.1233X_{t-5} - 0.3977X_{t-6} + Z_t \tag{9.44}$$

其中，X_t 代表第 t 时间点之预测需求量；$X_{t-k}(k=1,2,\cdots)$为需求订购量的历史数据；Z_t 为误差项。

2. 白噪声过程的验证

ARMA 过程的误差项假设为一白噪声过程，其亦属于平稳型序列，为支持平稳型时间序列模式建立的基础设定。因此，在模式构建完毕后须检验其残差是否满足白噪声过程。时间序列分布图虽能提供初步的平稳性检验，然而当序列中隐含过多噪声，使原序列波动情

形受到严重干扰，序列平稳性特质将不易被观察。本案例利用前述 Q 统计量来检验 $\{X_t\}$ $(t=0,\pm1,\pm2,\cdots)$ 序列的平稳性如下：

(1) 设立虚无假设与对立假设：

H_0：$\{X_t\}$ 为白噪声过程；

H_1：$\{X_t\}$ 非为白噪声过程。

(2) 选择显著水平：α 风险设为 0.05(会随数据量与风险函数而调整)。

(3) 找出对应的检定统计量：在 H_0 为真之下，Q 统计量的渐进分配为卡方分配，如式(9.45)所示：

$$Q = n\sum_{j=1}^{h}\hat{\rho}^2(j) \xrightarrow{d} \chi_h^2,\quad \hat{\rho}^2(j) = \mathrm{Cov}(t,t+j) \tag{9.45}$$

其中，n 为样本数，$\hat{\rho}^2(j)$ 为间隔 j 个时间单位的样本协方差函数。

(4) 决策法则的规定：拒绝域为 $Q<\chi_{h,\alpha}^2$ 或 $Q>\chi_{h,1-\alpha}^2$；p-value$<$0.05。

(5) 计算检定统计量：由剩余的数据波动计算样本协方差函数 $\hat{\rho}^2(j)$，并根据式(9.46)得检定统计值为

$$Q_0 = n\sum_{j=1}^{h}\hat{\rho}^2(j) = n\sum_{j=1}^{h}\mathrm{Cov}(t,t+j) \stackrel{\mathrm{def}}{=} \chi_0^2 \tag{9.46}$$

样本数据计算结果：p-value$=$0.527。

(6) 评估与结论：若 $\chi_{h,\alpha}^2<Q_0<\chi_{h,1-\alpha}^2$，则表示没有充分的证据显示所构建的 ARMA 过程的误差项不服从白噪声过程，因此可将所构建的模式应用于余波检测与未来值预测。若 $Q_0<\chi_{h,\alpha}^2$ 或 $Q_0>\chi_{h,1-\alpha}^2$，则表示有证据说明所构建的 ARMA 过程的误差项不服从白噪声过程，而需回到阶段一重新进行制程生命周期的配适。

因为 p-value$=$0.527$>$0.05，所以在 $\alpha=0.05$ 之下，不拒绝 H_0

即在显著水平为 0.05 下，没有充分的证据说明所构建的 ARMA 过程的误差项不服从白噪声过程。即 $ARMA(6,0)$满足基本假设。

分析结果发现并无显著异常的波动，因此以平稳型序列模式构建剩余波动的未来变化情形。本案例以 $ARMA(6,0)$为主要模式，构建出时间序列的预测模式，提供较准确的未来波动预测值。

图 9.22 为预测制程 A 的未来光罩需求量的曲线图，在考虑生命周期曲线与需求波动下，本案例根据制程 A 的历史顾客订单数据，通过建立的预测模式能对准确预测未来八季顾客对于光罩的需求订单量(如图 9.22 中右上方的灰色实线)，供管理者进行产能配置或需求满足管理。

9.9.4 案例小结

需求预测影响企业获利和成长，由于造成需求变动的因素是多维度且常有复杂的交互作用，导致需求预测问题大多以非结构化或半结构化的面貌呈现。因此，本案例以半结构化的生命周期曲线撷取需求变化的趋势，以及结构化的时间序列分析方法检测序列波动，并通过信度与效度验证结果，可以检验所提出的两阶段需求预测模式可提升预测准确率、降低预测误差变异。此外，亦可应用于相同背景下其他制程的光罩需求预测。

针对产品市场需求，作者(Chien *et al.*，2010)利用解释半导体产品需求的技术替代、价

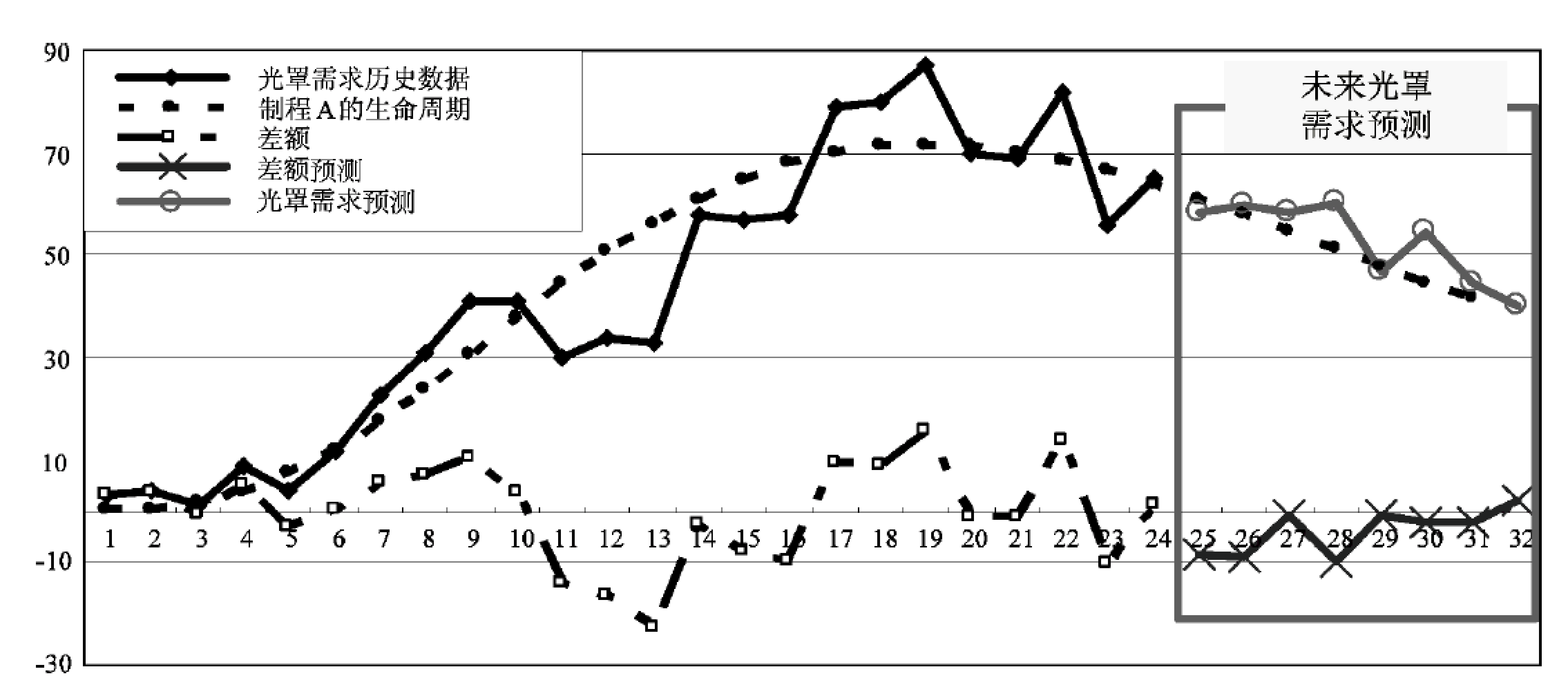


图 9.22 光罩需求于制程 A 的未来订单量预测

格、季节循环等因素，结合技术扩散模式，以发展需求预测模式，并导入半导体公司使用。

9.10 结论

多变量分析用以分析变量间的相关性与其背后数据结构，以作为预测或分类。本章介绍两种常见的多变量分析方法，每一种方法使用上都有其适用的数据形态与问题类型，研究人员必须决策采取最适合的分析方法。

虽然多变量分析法强调不只是分析数据，也要了解背后的因果关系(know why)，但实务上，研究人员所关心的现象，往往不止受到一项变量的影响。过去多变量分析在运算上相对较复杂、耗时且不易处理巨量数据，随着计算机运算能力增加，多变量分析的应用也越来越广泛。特别是在大数据分析时，往往是先找出数据呈现的样型和信息(know what)，尝试应用以创造价值或支持商业决策，再深入理解背后的因果关系。

时间序列数据能反映各类社会现象的发展过程和规律性，以预测未来的活动与发展趋势，有助于掌握动态且多变的商业和经济活动。目的是根据过去观察值所得到的规律、趋势或特殊样型以建立模式来预测未来区间，以供决策者参考。序列数据代表某时间间距下的一系列数列，其典型特征为该数据的反应值与时间相关。

时间序列的发展与回归统计分析息息相关。如时间序列中的自回归模式是利用时间序列中每一笔数据与前一期的数据进行回归计算，分析人员搜集及记录数据后，利用适当时间序列数据整理方法及统计分析计算工具定义出并估计回归模式的各参数，将参数代入方程式后进行分析预测。回归分析为静态预测，而时间序列分析可视为动态的预测。许多当下所发生的现象都是由前几期的现象或是由更早以前的事实逐步演化而来，而前期已发生的事实对于未来后续情况的演变往往具有或多或少的影响力，因此，根据过往的变动趋势和时间序列数据，即可预测未来可能发生的情况。企业或个人在从事任何涉及未来问题的决策都需运用到预测，唯有深入研究数列趋势的动态发展，才能得到准确可靠的预测。

问题与讨论

1. 线性回归与 CART 皆可拿来作为预测连续变量的模型。试比较两者的差异,并说明在哪些情况 CART 的表现会比线性回归好,反之亦然。

2. 假设 X 与 Y 的观测数据如下表所示,请回答下列各项问题:

(1) 请分别计算 X 与 Y 的样本平均与样本方差。

(2) 请计算 X 与 Y 的皮尔逊相关系数。

(3) 请问 X 与 Y 的相关性是属于低度相关、中度相关还是高度相关?

X	1	2	2	3	4	4	4	5
Y	5.2	5.9	6.9	8.6	13.1	11.3	10.6	12.4

3. 承上题,假设分析者欲使用最小二乘估计法建构 X 对 Y 的回归模式 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$,请回答下列问题:

(1) 请绘制 X-Y 之散布图。

(2) 请计算回归式中,$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的数值。

(3) 请将所估计的回归线绘制于(1)的散布图中。

(4) 请对各笔数据计算其预测值与残差值。

(5) 请计算此回归式的 SSE、SSR、SST 与 $\hat{\sigma}^2$。

(6) 请计算此回归式的 R^2 与 R_a^2。

4. 假设 X 与 Y 的观测数据如下表所示,假设使用最小二乘估计法构建 X 对 Y 的回归模式 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$,请回答下列问题:

(1) 请计算回归式中,$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的数值。

(2) 请对所构建的模式进行方差分析,并根据 F 统计量说明此模式在统计上是否显著。

(3) 请绘制 X-Y 的散布图,并根据(1)所估计的参数将回归线绘制于散布图中。

(4) 请根据以上结果,论述此例中的 X 与 Y 是否有关系。

X	1	2	2	3	4	4	4	5
Y	12.4	4	3.5	−5.8	0.5	0.4	1.4	11.7

5. 下表为针对工作压力所进行的抽样调查结果,其中压力字段表示受访者自觉工作压力过大的情况。请回答下列问题:

(1) 请问在所有受访者当中,工作压力过大的胜算为多少?

(2) 请分别计算男性与女性受访者工作压力过大的胜算。

(3) 请分别计算四种血型受访者工作压力过大的胜算。

(4) 请计算 O 型血男性受访者工作压力过大的胜算。

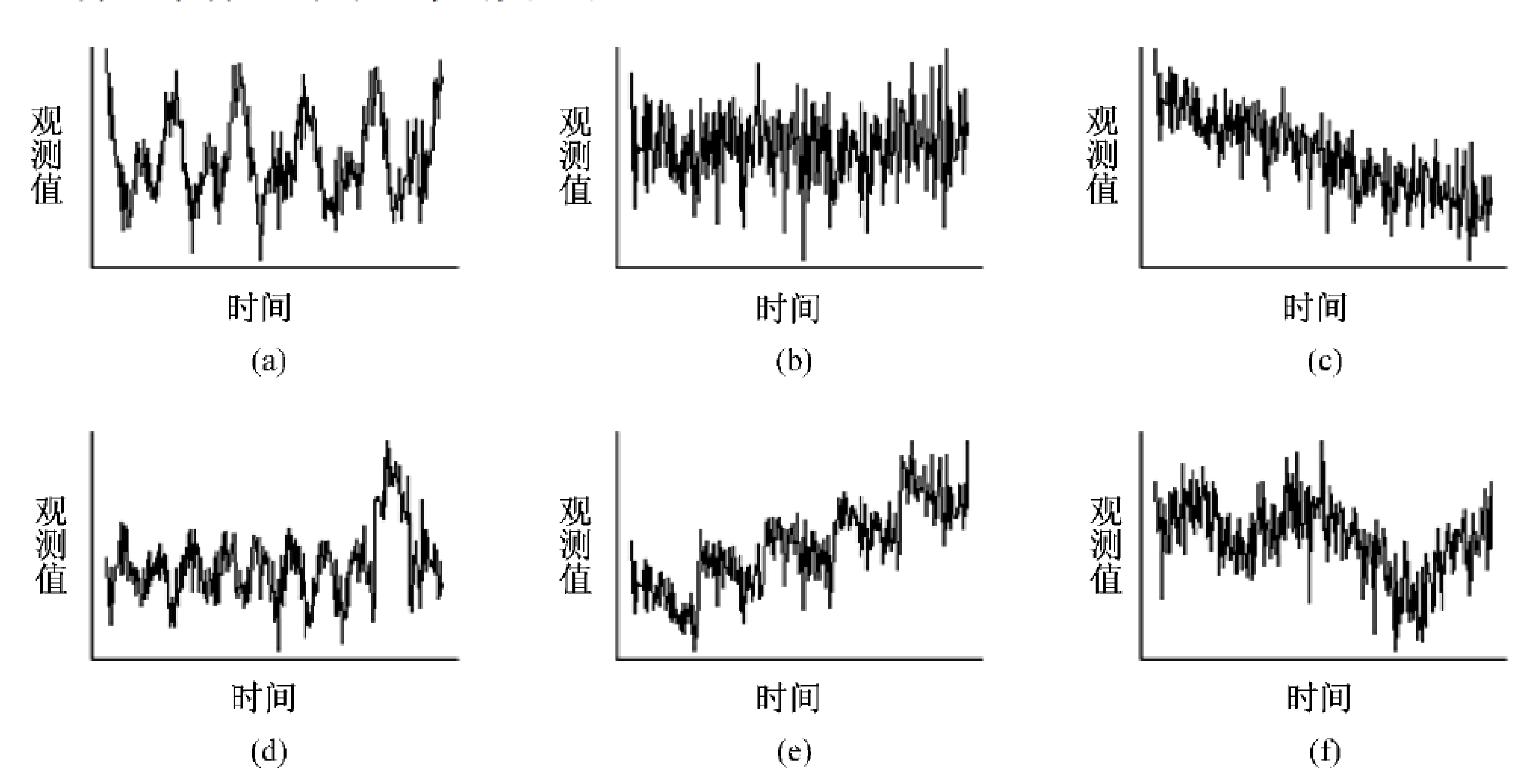
编号	性别	血型	压力	编号	性别	血型	压力
1	男	O	有	11	男	O	无
2	男	A	有	12	女	B	无
3	男	A	有	13	女	AB	有
4	女	B	无	14	男	O	有
5	女	O	无	15	男	O	无
6	男	O	无	16	女	O	无
7	男	A	有	17	男	A	无
8	女	AB	无	18	女	B	有
9	女	O	无	19	女	AB	无
10	男	B	无	20	男	O	有

6. 假设事件 A 发生的概率可写成 $P(A|X=x)=e^{2+3x}/(1+e^{2+3x})$，请回答下列问题：

(1) 请计算当 $X=-1,0,1$ 的时候，A 发生的概率分别为多少？

(2) 请问当 X 为多少的时候，A 发生的概率会为 0.6？

7. 请判断下列序列(a)～(f)分别属于：(1)平稳型、(2)无定向型、(3)趋势型、(4)季节型、(5)介入事件型的哪些序列类型？



8. 假设序列 $\{a_t\}_{t=1}^{20}=(-0.6, 0.4, -0.5, 0.4, -0.3, 0.3, 0.2, 0.3, 1.0, 0.8, -0.8, -1.6, -2.1, 0.2, 0.9, -1.0, 1.4, -0.4, -0.1, 1.3)$，请回答下列问题：

(1) 请绘制序列 $\{a_t\}_{t=1}^{20}$ 的趋势图。

(2) 请计算序列 $\{a_t\}_{t=1}^{20}$ 的 1 阶自相关系数。

(3) 请使用移动平均过程计算 $Z_t=a_t+\sum_{k=1}^{4}a_{t-k}(t=5,\cdots,20)$，并将其绘制于(1) 所绘制的趋势图中。

(4) 请计算 $\{Z_t\}_{t=5}^{20}$ 的 1 阶自相关系数。

9. 假设$\{Z_t\}_{t=1}^{15}$=(−0.7，−2.4，−1.1，−0.1，−0.6，0.5，1.0，1.1，2.3，2.6，2.1，1.6，0.2，0.3，0.6)为一时间序列的观测值，请回答下列问题：

(1) 请绘制$\{Z_t\}_{t=1}^{15}$的趋势图。

(2) 请绘制$\{Z_t\}_{t=1}^{14}$对$\{Z_t\}_{t=2}^{15}$的散布图。

(3) 假设$X=\{Z_t\}_{t=1}^{14}$、$Y=\{Z_t\}_{t=2}^{15}$，请利用最小二乘估计法对X与Y构建回归模式$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$，并检定$X$与$Y$之间的线性关系是否显著。

10. 假设$\{Z_t\}_{t=1}^{20}$=(0，0.16，0.23，0.53，0.17，0.58，0.55，0.13，−0.09，−0.06，−0.05，−0.49，−0.24，−0.3，−0.21，−0.05，0.26，0.34，0.5，0.24)为一时间序列的观测值，请回答下列问题：

(1) 请估计此序列的自相关系数$\hat{\rho}_k(k=1,2,3)$。

(2) 请估计此序列的偏自相关函数$\hat{\rho}_{22}$。

(3) 请使用$AR(1)$模式配适此时间序列数据，并写出各项参数的显著性。

(4) 请使用$AR(2)$模式配适此时间序列数据，并写出各项参数的显著性。

(5) 由(2)～(4)的结果，请问$AR(1)$与$AR(2)$中，何者较适合用来解释此数据集？

11. 试产生下列各种样型的时间序列：平稳型、无定向型、趋势型、季节型、介入事件型，并画出以下图形：

(1) 序列图。

(2) 自相关函数。

(3) 偏自相关函数。

(4) 一阶差分后的序列图。

12. 令$X_t=\theta X_{t-1}+Z_t$为一自$AR(1)$过程，其中$\{Z_t\}\sim WN(0,\sigma^2)$。试针对$\theta$探讨$\{X_t\}$的平稳性。

13. 令$X_t=Z_t+0.5Z_{t-1}$为一移动平均过程，其中$\{Z_t\}\sim WN(0,1)$，请试着回答下列问题：

(1) 计算$\{X_t\}$过程的自相关函数。

(2) 推论$Y=X_3-X_4+X_5$的分布状况。

(3) 推论(X_3,X_5)的联合概率密度函数。

(4) 在已知$X_3=1$的情况下，试推论X_4的分布状况；反之，在已知$X_4=1$的情况下，试推论X_3的分布状况。

第 10 章

集成学习与支持向量机

10.1 集成学习

集成学习方法(ensemble learning method)是为了改善分类预测准确率的一种学习算法,以决策树分析为例,集成学习方法提供了如何不修剪决策树分支而提高测试资料的预测准确性。集成学习算法的计算过程是构建一组由多个分类结果组合而成的分类模型,再经由多个分类结果的投票(voting),用以预测未知数据的卷标类别或数值,最后的分类模型将取决于个别模型分类结果与对应的权重大小。举例而言,个别分类模型的结果如同病人的疾病诊断往往仅由一位医生依其病征决定,因此,该位医生是否提供正确的诊断就变得很重要。集成学习的概念上主要则是将同一位病人借由不同医生的诊断进行综合判断,不同医生的诊断效力相同,如果同一种诊断结果在不同医生间重复出现,则该诊断可视为该病人所发生的疾病。

相较于个别算法所建立的分类模型,集成学习算法已被许多学者证实出其表现具有显著的改进。集成学习算法主要有两种方法:第一种方法是产生多样的不同模式,基于不同模式的预测准确性不尽相同,如果在多数不同模式间所预测的结果具有大部分的一致性,相对于个别模式而言,可降低其发生错误的状况;第二种方法是改变不同模式在预测结果的权重大小,也就是提高预测准确性较佳的模型权重,并降低预测准确性不佳的模型权重,进而整合不同模型的权重以产生更接近实际结果的分类模型。

Bagging 与 Boosting 为两种常见的集成学习方法,此两种方法的演算机制均为先选定一学习理论做基础运算,之后辅以不同的分类算法与训练样本组进而找寻最佳分类模型。不论是以重复抽取的方式找出新的样本组合,或是调整权重产生新样本值,Bagging 与 Boosting 集成学习方法都可有效提高分类准确性。

10.1.1 Bagging

Bagging 为拔靴整合(bootstrap aggregating)的缩写(Breiman, 1996),在 bootstrap 阶段,Bagging 学习算法结合的目的在于产生具有多样性的训练数据子集合。如图 10.1 所示,Bagging 产生的方式可从原先训练数据组中重复建立取样的训练数据子集合,也就是说,给定原始的一组训练样本组 S,其中包含 m 个样本数,接着重复抽样并重组为另一组同样具有 m 个样本数的新样本组产生新的训练样本组 S'。在整合阶段,对于分类问题则用投票的方式决定预测类别,对于回归问题则利用平均数作为预测值。Bagging 可避免单一模型的分类发生高度变异的情况,与仅有一个分类模型相比,不仅具有较高的正确率,受到噪声数据的影响程度也较小。

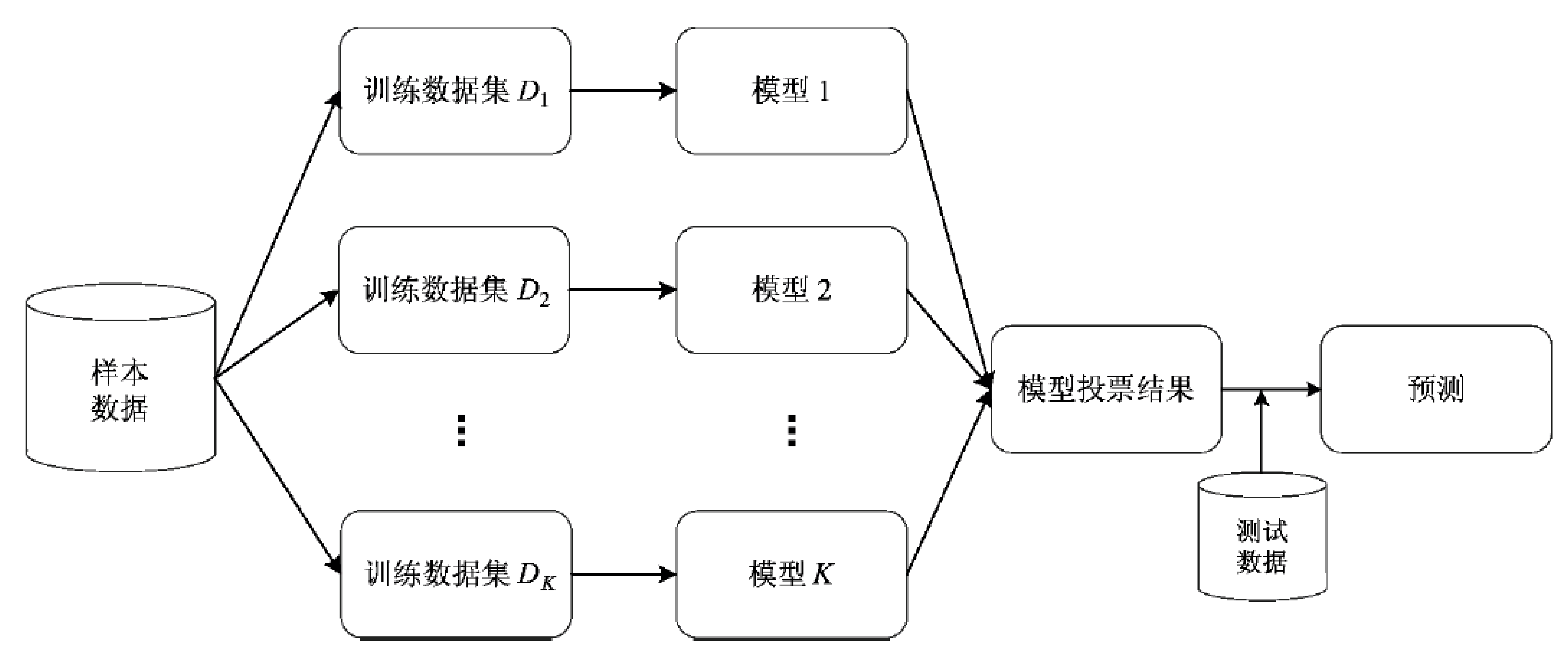


图 10.1　Bagging 学习算法

确保个别分类模型的多样性是 Bagging 算法提高准确性的关键，较直接的方式为选择不同的输入特征子集合与随机性(randomness)的特征。布赖曼(Breiman，2001)整合 Bagging 与 random subspace(Ho，1998)提出随机森林(random forests)，不同于决策树每次仅产生一棵树作为分类模型，随机森林则是利用森林作为最后的分类模型，随机森林以 CART 决策树算法作为长树的方法，在每个分支节点随机选择数个属性作为分支变量，改变分类模型的预测变量以产生不同的模型。

给定一组样本数据 $D=(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)$，每笔样本数据 $\boldsymbol{x}_i$ 有 M 个属性 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{iM})$，样本数据 y_i 为目标属性。随机森林算法说明如下。

阶段 1：训练模型

(1) 利用拔靴法从样本数据集 D 中选取 N 笔数据形成另一个训练数据集 D_j，$j=1,2,\cdots,K$。

(2) 在每个节点上随机选取 m 个属性作为决策树长树候选属性($m<M$)，根据所选择的 m 个属性计算其最佳树枝生长结果，每一颗决策树将不断分支，直到所有候选属性均无法满足分支条件为止。

阶段 2：预测

(1) 输入新的样本数据 $\boldsymbol{x}'$ 分别至 K 个决策树。

(2) 如果目标属性为类别属性，则新样本数据 $\boldsymbol{x}'$ 的预测类别为 K 个决策树中的多数类别。

(3) 如果目标属性为连续属性，则新样本数据 $\boldsymbol{x}'$ 的预测值为 K 个决策树模型的预测平均值。

随机森林可快速地处理大量且高维度的数据，由于在每次分支时仅选用部分的属性数据，因此对于大量且高维度的数据能有很好的计算效率，而随机森林重复抽取训练数据的做法，也使得其分析结果较不易受到噪声与异常值的影响。

10.1.2　Boosting

集成学习的另一种方式为采用加总模式(additive model)预测未知数据的类别，其中该

加总模式由许多不同的分类模型所构成，个别分类模型所产生的误差即为权重大小，误差越大则在加总模式的权重越小，误差越小则在加总模式的权重越大。

AdaBoost 是 Boosting 学习算法中最著名的算法，AdaBoost 为 adaptive boosting 的缩写，不同于 Bagging 每次学习过程中会不断改变训练数据的组成，AdaBoost 的训练数据均为同一组。AdaBoost 目的为，在学习过程时，借由不断调整分类数据的权重值以尽可能地降低训练样本的分类错误，对于分类错误的数据会给予权重的调整，使得在下一次学习上得以改善其分类结果；反之，当该样本数据分类结果正确时，则会降低该样本数据的权重值，最后产生 K 个分类模型，并依据每次的学习所产生的分类结果的权重进行加权，得到最后的分类模型。

给定一组二元分类的样本数据 $S=(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)$，每笔样本数据 $\boldsymbol{x}_i$ 有 M 个属性 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{iM})$，样本数据 $y_i\in\{-1,+1\}$，D_j 代表第 j 次的训练迭代，$D_j(i)$代表第 j 次的训练模型中第 i 笔数据的权重，Adaboost 算法说明如下。

(1) 设定 $D_1(i)=1/N(i=1,2,\cdots,N)$。

(2) 产生分类模型 $h_j(j=1,2,\cdots,K)$，并计算在分类模型 h_j 下的分类错误率 ε_j，

$$\varepsilon_j=\sum_{i=1,h_j(x_i)\neq y_i}^{N}D_j(i) \tag{10.1}$$

其中，$h_j(x_i)\neq y_i$ 代表当模型 $h_j(x_i)$的预测结果与实际类别 y_i 不同，此外，如果 $\varepsilon_j>0.5$，则重新回到步骤(1)。

(3) 计算该分类模型 h_j 的重要性 α_j，$\alpha_j=\frac{1}{2}\ln\left(\frac{1-\varepsilon_j}{\varepsilon_j}\right)$。 (10.2)

(4) 调整 S 中各样本的权重 $D_{j+1}(i)$作为下一次学习样本数据权重

$$D_{j+1}(i)=\frac{D_j(i)\exp(-\alpha_j y_i h_j(x_i))}{Z_j} \tag{10.3}$$

其中，Z_j 为归一化因子，为确保所有 $D_{j+1}(i)$的和为 1，$Z_j=2\ [\varepsilon_t(1-\varepsilon_t)]^{1/2}$

(5) 学习 K 次后依据 K 个分类模型 h_j，得到最终分类模型 H：

$$H(x)=\text{sgn}\left(\sum_{j=1}^{K}\alpha_j h_j(x)\right) \tag{10.4}$$

图 10.2、图 10.3 为说明 AdaBoost 算法的计算过程，共有 10 笔训练样本，其中，5 笔为 +1(标记为三角形)，5 笔资料为 −1(标记为圆形)，设定学习循环为 $K=3$。

表 10.1 说明其 3 次循环的计算结果：在第 1 次循环中，给定所有的训练样本数据权重为 1/10，根据分类模型 h_1的结果计算其分类错误率，发现样本点 6、样本点 7、样本点 9 分类错误，因此分类错误率 ε_1 为 0.30，根据式(10.2)计算分类模型 1 的权重 $\alpha_1=\frac{1}{2}\ln\left(\frac{1-0.30}{0.30}\right)\approx 0.424$，接着根据式(10.3)与归一化因子 Z_1，$Z_1=2\sqrt{0.3\times 0.7}\approx 0.917$，增大先前分类错误的 3 个样本点的权重并

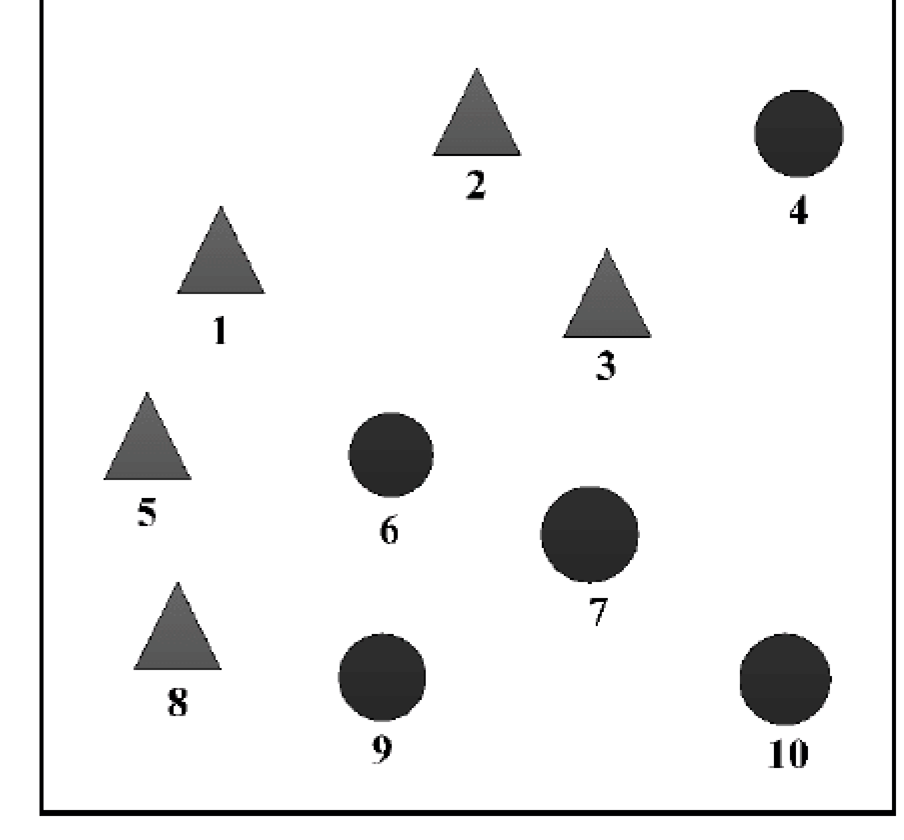


图 10.2 AdaBoost 计算范例

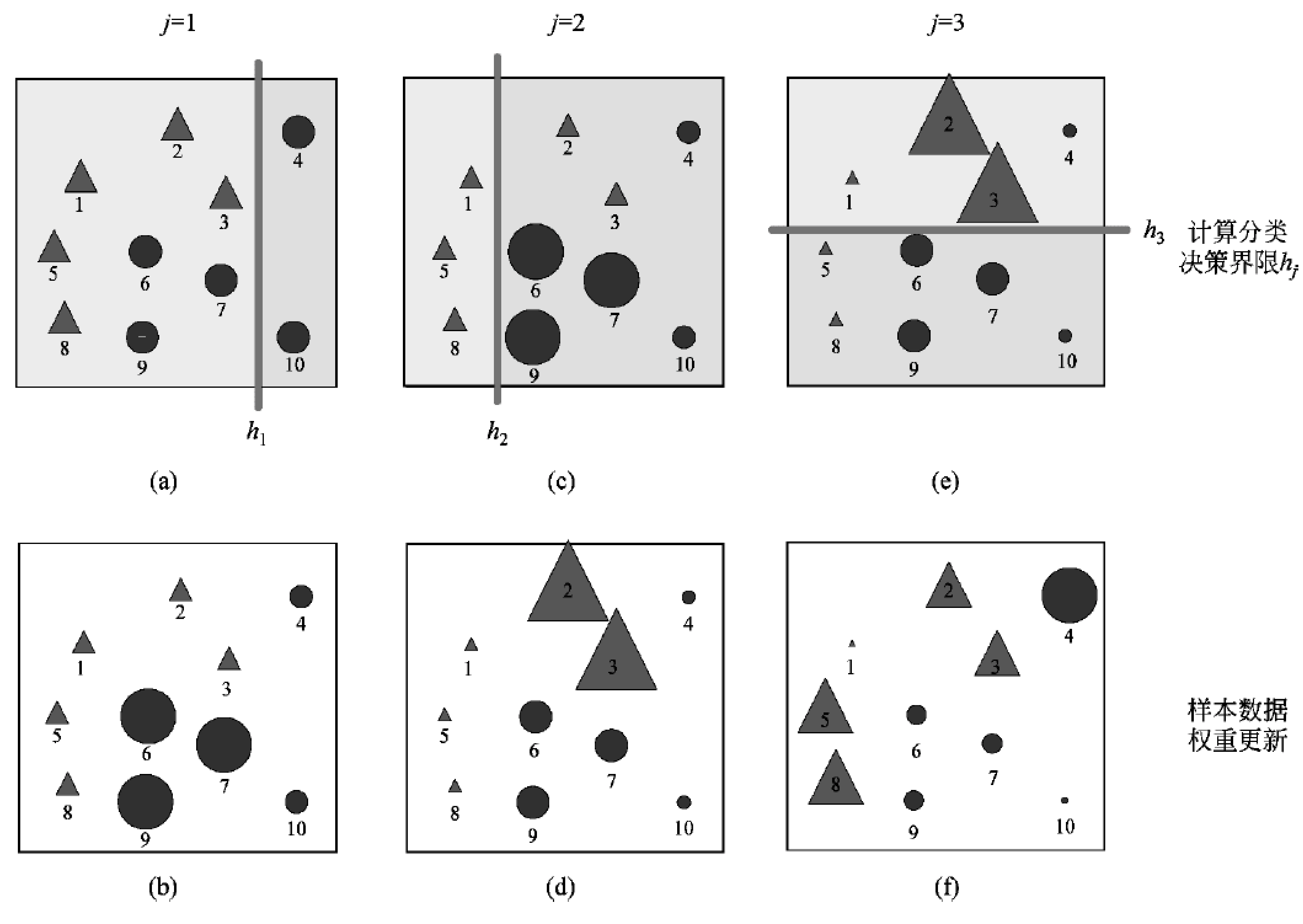


图 10.3　AdaBoost 范例计算过程(K=3)

降低其他 7 个样本点的权重 $D_2(i)$，使得分类错误的 3 个样本点在下一次分类中尽可能地被考虑；在第 2 次循环中，根据 $D_2(i)$ 的权重，在分类模型 h_2 下，发现样本点 2、样本点 3 为分类错误的数据，接着计算其分类错误率 $\varepsilon_2 \approx 0.143$，计算分类模型 2 的权重为 $\alpha_2 = \frac{1}{2}\ln\left(\frac{1-0.143}{0.143}\right) \approx 0.896$，最后根据式(10.3)与归一化因子 $Z_2 \approx 0.700$ 更新每个训练样本的权重 $D_3(i)$，第 3 次循环依据分类模型 h_3 的结果得到 $\varepsilon_3 \approx 0.125$ 以及 $\alpha_3 \approx 0.973$。

表 10.1　AdaBoost 范例计算结果

	1	2	3	4	5	6	7	8	9	10
$D_1(i)$	0.10	0.10	0.10	0.10	0.10	**0.10**	**0.10**	0.10	**0.10**	0.10
$e^{-\alpha_1 y_i h_1(x_i)}$	0.65	0.65	0.65	0.65	0.65	1.53	1.53	0.65	1.53	0.65
$D_1(i)e^{-\alpha_1 y_i h_1(x_i)}$	0.07	0.07	0.07	0.07	0.07	0.15	0.15	0.07	0.15	0.07
$\varepsilon_1=0.30$，$\alpha_1\approx0.424$，$Z_1\approx0.917$										
$D_2(i)$	0.07	**0.07**	**0.07**	0.07	0.07	0.17	0.17	0.07	0.17	0.07
$e^{-\alpha_2 y_i h_2(x_i)}$	0.41	2.45	2.45	0.41	0.41	0.41	0.41	0.41	0.41	0.41
$D_2(i)e^{-\alpha_2 y_i h_2(x_i)}$	0.03	0.17	0.17	0.03	0.03	0.07	0.07	0.03	0.07	0.03
$\varepsilon_2\approx0.143$，$\alpha_2\approx0.896$，$Z_2\approx0.700$										

续表

	1	2	3	4	5	6	7	8	9	10
$D_3(i)$	0.04	0.25	0.25	**0.04**	**0.04**	0.10	0.10	**0.04**	0.10	0.04
$e^{-\alpha_3 y_i h_3(x_i)}$	0.38	0.38	0.38	2.65	2.65	0.38	0.38	2.65	0.38	0.38
$D_3(i)e^{-\alpha_3 y_i h_3(x_i)}$	0.02	0.09	0.09	0.11	0.11	0.04	0.04	0.11	0.04	0.02
$\varepsilon_3 \approx 0.125$，$\alpha_3 \approx 0.973$，$Z_3 \approx 0.661$										
$D_4(i)$	0.02	0.14	0.14	0.17	0.17	0.06	0.06	0.17	0.06	0.02

样本数据点 x 的类别为根据各分类模型 h_1、h_2、h_3 与其权重 $\alpha_1 \approx 0.424$、$\alpha_2 \approx 0.896$、$\alpha_3 \approx 0.973$，加权计算后得到整合分类模型 $H(x)$，如图 10.4。

$$H(x) = \text{sgn}(\alpha_1 h_1(x) + \alpha_2 h_2(x) + \alpha_3 h_3(x))$$

$H = 0.424$ [图] $+ 0.896$ [图] $+ 0.973$ [图]

$=$ [图]

图 10.4 AdaBoost 整合分类结果

以样本点 2 为例，$H(2)=\text{sgn}(0.424-0.896+0.973)=\text{sgn}(0.501)=+1$；

以样本点 9 为例，$H(9)=\text{sgn}(0.424-0.896-0.973)=\text{sgn}(-1.445)=-1$。

10.2 支持向量机

支持向量机(support vector machine，SVM)是一种监督式学习的方法，主要可用于分类(classification)或回归(regression)类型的问题(Cortes & Vapnik，1995)。SVM 算法是将原始数据特征转换至另一个高维度，并基于构建一个或多个超平面(hyperplane)，使得训练数据中不同类别的数据得以尽可能地分开，同时该超平面需尽可能地远离各类别中最靠近超平面的数据点(Vapnik，1995)。超平面即为分类边界，超平面与各类别最近的训练数据点的距离为边缘(margin)。因此，SVM 的学习目的在于找到具有最大边缘的超平面(maximum marginal hyperplane)。以图 10.5 二元分类为例，共有类别 A 与类别 B，要找到一超平面得以将两个类别正确分开，其中超平面 1 与超平面 2 均可将两个类别的数据正确地划分，但超平面 1 因为拥有较大的边缘，因此超平面 1 的分类效果优于超平面 2。

10.2.1 可区分情况(separable case)

给定在 N 维度空间的训练数据 $D=\{\boldsymbol{x}_i, y_i \mid \boldsymbol{x}_i \in \mathbb{R}^N, y_i \in \{-1,1\}, i=1,2,\cdots,m\}$，$y_i$ 代表资料 $\boldsymbol{x}_i$ 所属的类别，标记为 -1 或是 1，假设期望找到一个超平面得以尽可能地分开两个

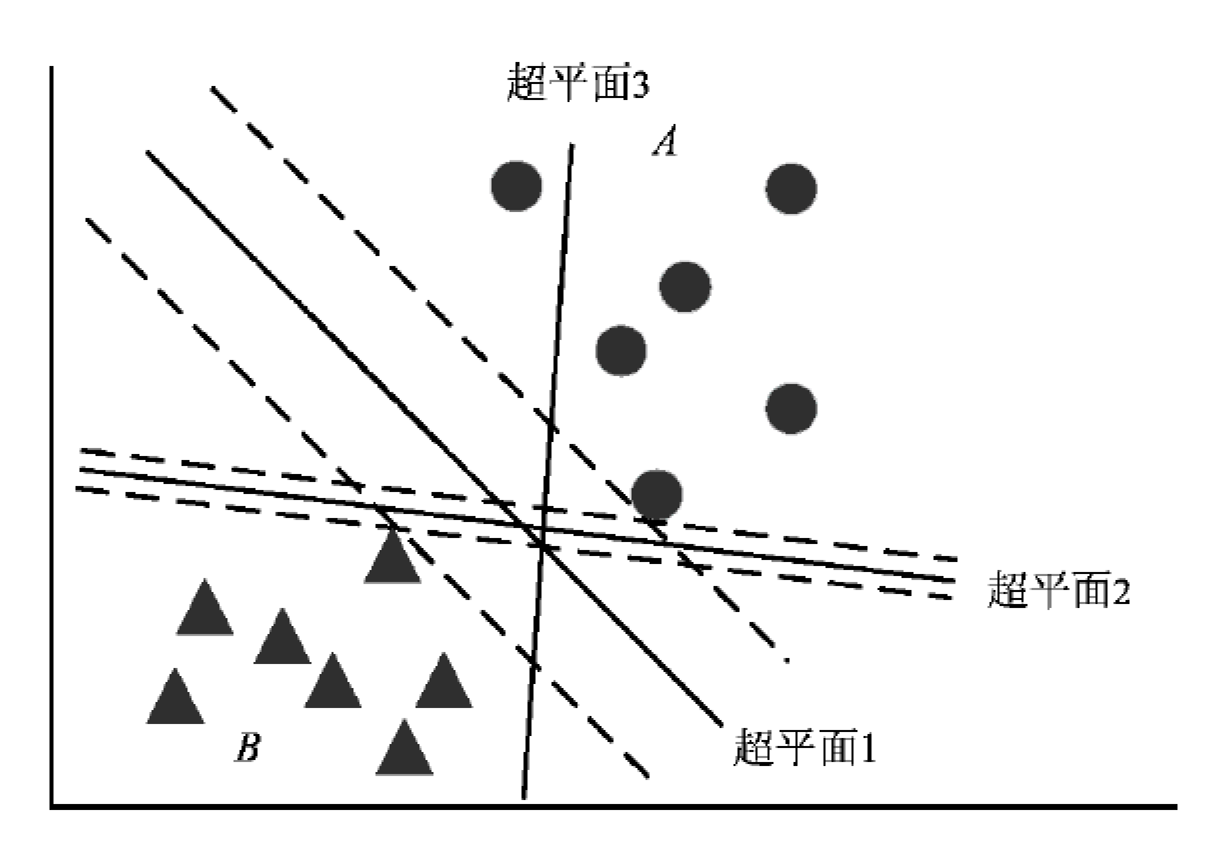


图 10.5　超平面与支持向量

类别,同时所有坐落在该超平面的数据点 $\boldsymbol{x}$ 均满足:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \tag{10.5}$$

其中,$\boldsymbol{w}$ 为垂直于超平面的非零(non-zero)向量 $\boldsymbol{w} \in \mathbb{R}^N$,$b$ 为位移量,$b \in \mathbb{R}$。若训练数据为线性可分,通过调整 $\boldsymbol{w}$ 与 b 可以找到两个临界超平面(marginal hyperplane)H_1 与 H_2,如图 10.6 所示,两个超平面可定义如下:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 1 \tag{10.6}$$

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = -1 \tag{10.7}$$

其中,±1 为常数。因此,可得到两个超平面的距离为 $2/\|\boldsymbol{w}\|$,为了使所有数据点均落在两个超平面之外(两个超平面之间没有任何的样本点),所有数据点 $\boldsymbol{x}_i$ 需满足以下两个不等式其中之一:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b \geqslant 1, \quad y_i = 1 \tag{10.8}$$

$$\boldsymbol{w} \cdot \boldsymbol{x} + b \leqslant -1, \quad y_i = -1 \tag{10.9}$$

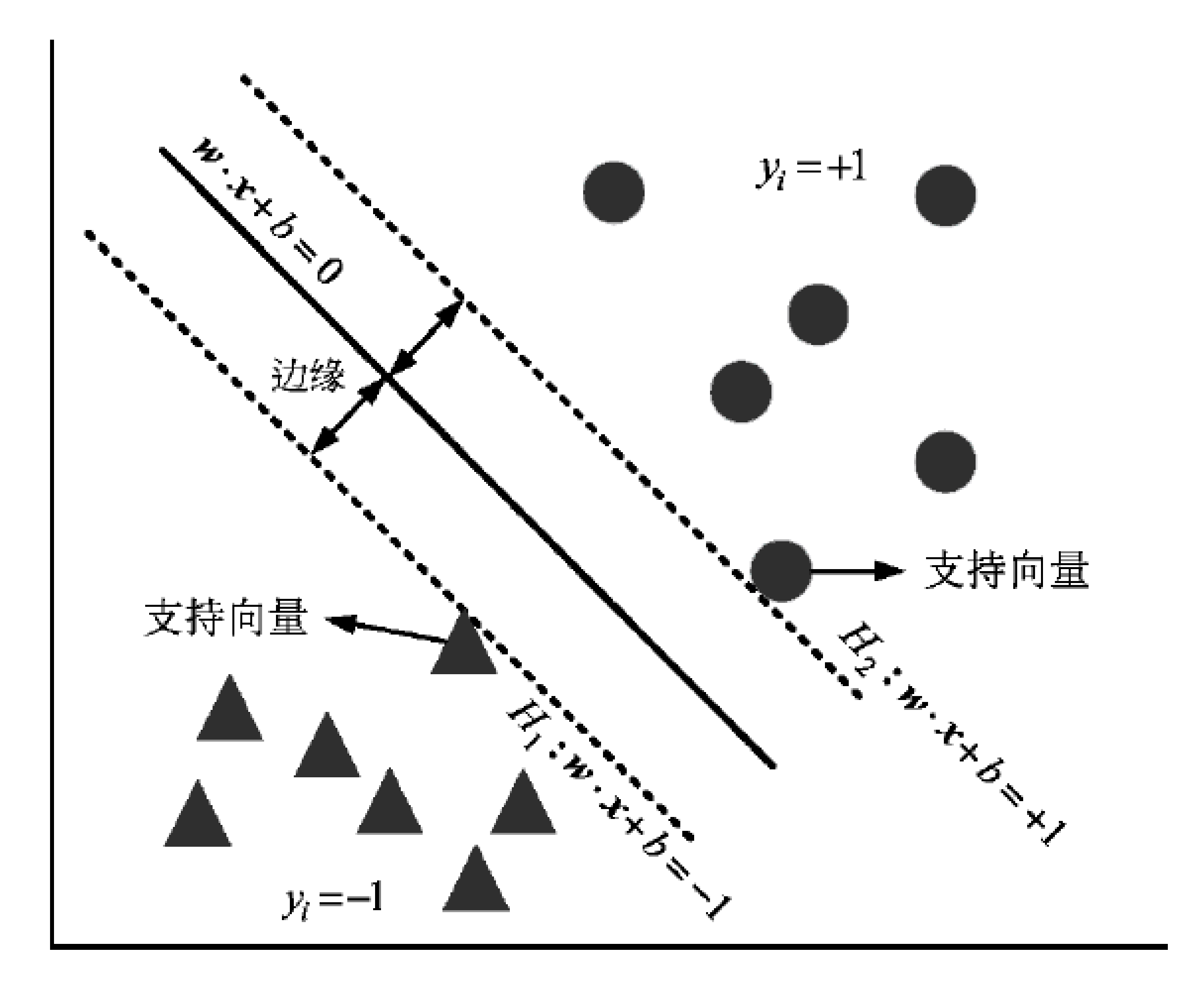


图 10.6　超平面与支持向量

式(10.8)与式(10.9)可合并为

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x} + b) \geqslant 1 \tag{10.10}$$

因此,在线性可分割的案例,SVM 最佳超平面可表示为二次规划(quadratic programming)的优化问题,表示如下:

$$\min_{w,b} \frac{1}{2} \| \boldsymbol{w} \|^2$$

$$\text{subject to } y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geqslant 1, \quad i = 1,2,\cdots,m \tag{10.11}$$

为了解以上优化问题,可利用非负的拉格朗日乘数(Lagrange multiplier)α_i,$\alpha_i \geqslant 0$,得到拉格朗日函数如式(10.12):

$$L(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2} \| \boldsymbol{w} \|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1] \tag{10.12}$$

再利用二次规划求解技术,分别对 $\boldsymbol{w}$ 与 b 偏微分,可求得一最佳解使得

$$\nabla_w L = \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i = \boldsymbol{0} \Rightarrow \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{10.13}$$

$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = \boldsymbol{0} \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{10.14}$$

$$\alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1] = 0, \quad i = 1,2,\cdots,m \tag{10.15}$$

根据式(10.13)可得知,权重向量 $\boldsymbol{w}$ 为训练数据集所产生的线性组合,而仅有少数的数据 $\boldsymbol{x}_i$ 会实际对目前函数有所影响,也就是仅有少数的 α_i 会大于 0,这些资料又称为支持向量(support vector)。式(10.15)也可以确保支持向量必定落在临界超平面,如果 $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) = 1$,则 $\alpha_i \neq 0$,如果 $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) > 1$,$\alpha_i = 0$。

因为所有支持向量 $\boldsymbol{x}_i^{sv}$ 使得 $y_i = \boldsymbol{w} \cdot \boldsymbol{x}_i^{sv} + b$,因此 SVM 系数 b 可定义为

$$b = \boldsymbol{w} \cdot \boldsymbol{x}_i^{sv} - y_i = \sum_{j=1}^{m} \alpha_j y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}_i^{sv}) \tag{10.16}$$

当有测试数据集 $\boldsymbol{x}^t$,可利用以下最大超平面方程式结果判断:

$$h(\boldsymbol{x}^t) = \operatorname{sgn}(\boldsymbol{w} \cdot \boldsymbol{x}^t + b) = \operatorname{sgn}\left(\sum_{j=1}^{m} \alpha_j y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}^t) + b \right) \tag{10.17}$$

当 $h(\boldsymbol{x}^t) = 1$,表示预测 $\boldsymbol{x}^t$ 类别为 $+1$,如果 $h(\boldsymbol{x}^t) = -1$,则预测 $\boldsymbol{x}^t$ 类别为 -1。

10.2.2 不可分状况(non-separable case)

当训练数据集合中不是线性可分割时,也就是无法找到一个超平面 $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ 得以将所有的训练数据 $\boldsymbol{x}_i$ 正确地区分,则限制式(10.10)可加入一个松弛变量(slack variable)ξ_i,$\xi_i \geqslant 0$,使得不等式成立,如式(10.18):

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geqslant 1 - \xi_i \tag{10.18}$$

松弛变量 ξ_i 用以表示训练数据集中违反不等式(10.10)的距离,如图 10.7 所示。因此,当 $\xi_i > 0$,表示该资料无法正确借由超平面 $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ 所分类,扣除无法被正确分类的训练数据点所形成的超平面,其边缘为 $1/\| \boldsymbol{w} \|$,相对于可分割的(separable)例子又称为柔性边缘(soft margin)。

在不可分割例子中,SVM 目的除尽可能地找到最大化边缘超平面外,也要最小化无法正确区分训练数据的误差,因此,可在原优化问题的目标式中加入参数 C,并在目标式中加入无法正确区分所造成的误差,$C > 0$:

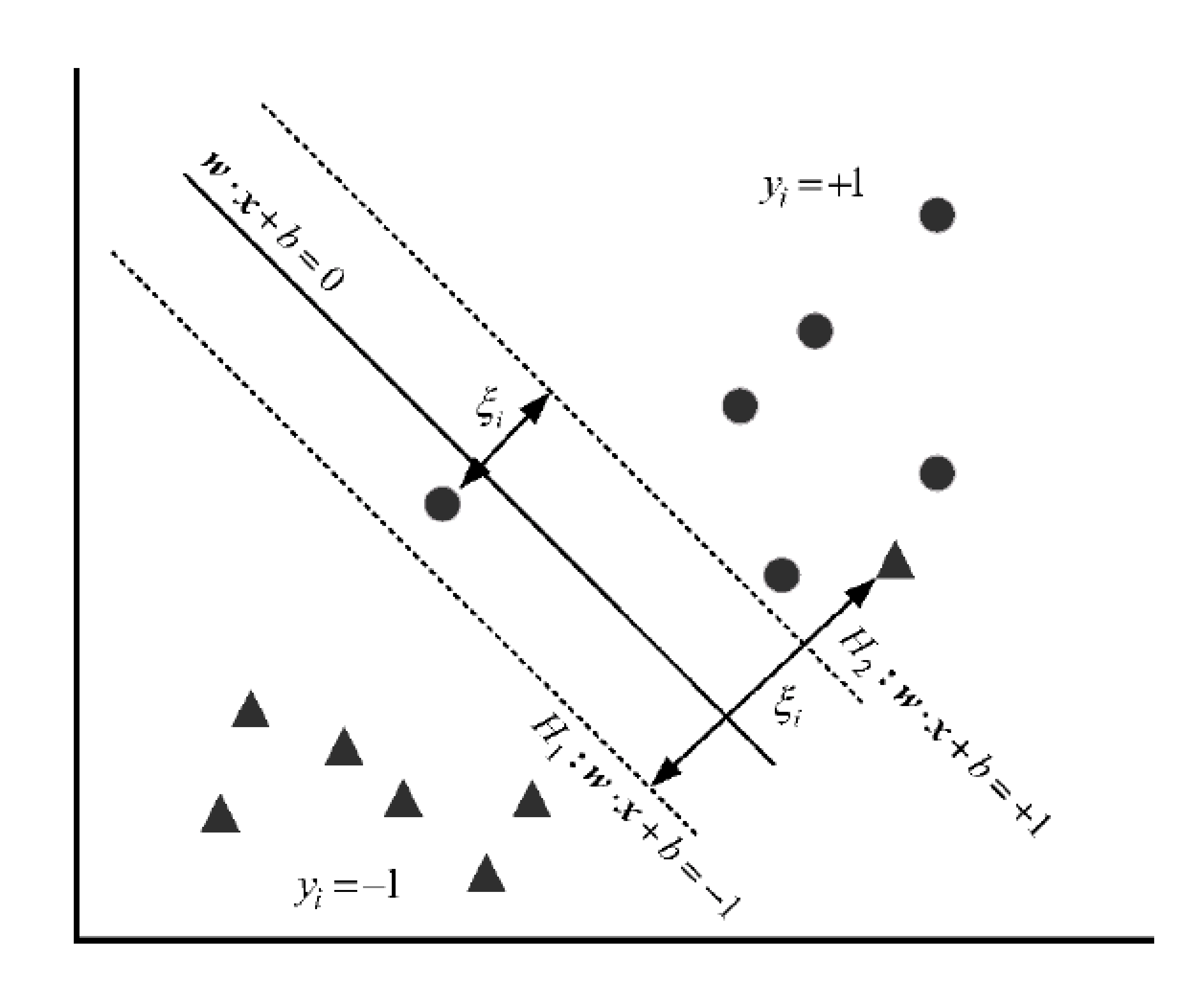


图 10.7　不可分割范例

（在超平面 $y_i(w\cdot x_i+b)\geqslant 1$ 下，类别+1 与−1 分别有一笔数据点为分类错误）

$$\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t. } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1-\xi_i,\quad i=1,2,\cdots,m \tag{10.19}$$

其中，$\xi=(\xi_1,\xi_2,\cdots,\xi_m)^{\mathrm{T}}$。参数 C 的决定可利用 k-folds 交叉验证决定最佳的参数值。

再利用拉格朗日(Lagrangian)转换与 KKT 条件进行求解，给定 α_i 与 β_i 分别对应 m 条限制式以及 m 个非负的松弛变量限制式，$\alpha_i\geqslant 0$、$\beta_i\geqslant 0$，可得到拉格朗日函数如式(10.20)：

$$L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1+\xi_i]-\sum_{i=1}^{m}\beta_i\xi_i \tag{10.20}$$

令拉格朗日函数对 $\boldsymbol{w}$、b、ξ_i 偏微分后为 0，加上其充分条件如下：

$$\nabla_w L=\boldsymbol{w}-\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i=\boldsymbol{0}\Rightarrow\boldsymbol{w}=\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i \tag{10.21}$$

$$\nabla_b L=-\sum_{i=1}^{m}\alpha_i y_i=0\Rightarrow\sum_{i=1}^{m}\alpha_i y_i=0 \tag{10.22}$$

$$\nabla_{\xi_i} L=C-\alpha_i-\beta_i=0\Rightarrow\alpha_i+\beta_i=C \tag{10.23}$$

$$\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1+\xi_i]=0,\quad i=1,2,\cdots,m \tag{10.24}$$

$$\beta_i\xi_i=0,\quad i=1,2,\cdots,m \tag{10.25}$$

如同可分割的例子中，根据式(10.21)得知权重向量 $\boldsymbol{w}$ 为训练数据集所产生的线性组合，在线性不可分的例子当中，当 $\alpha_i\neq 0$ 时，$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)=1-\xi_i$，若 $\xi_i=0$，则代表对应的 $\boldsymbol{x}_i$ 落在超平面上，因此 $\boldsymbol{x}_i$ 为支持向量；若 $\xi_i\neq 0$，根据式(10.25)，则 $\beta_i=0$，因此，该 $\boldsymbol{x}_i$ 为无法正确区分的训练数据。

10.2.3　非线性分类

当数据为无法以线性区分时，线性 SVM 无法找到适合的解，此时可改用非线性转换函

数 Φ 将原输入样本空间(sample space) X 映射至一个高维度的特征空间 H(feature space)中,也称为 Hilbert 空间,找到一个非线性(non-linear)的决策界线(decision boundary),使得原本样本空间中非线性可分割的问题转换为特征空间中线性可分割的问题。在计算最大边缘的超平面上也会需要大量的高维度内积计算,因此 SVM 在处理非线性转换 $\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j)$上采用核函数(kernel function) $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$表示,定义如下:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j) \tag{10.26}$$

在计算核函数上远比直接计算 $\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j)$来得有效率,甚至可以不需要知道非线性转换函数的正确方程式。常用的核函数有以下三种(表 10.2)。

表 10.2 SVM 核函数

核 函 数	数 学 式
多项式函数 (polynomial function)	$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i,\boldsymbol{x}_j+c)^d, c>0, d\in\mathbb{N}$ c 为常数,d 为多项式的次方项,例如二次多项式则 $d=2$
高斯径向基函数 (Gaussian radial basis function)	$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\Vert\boldsymbol{x}_i-\boldsymbol{x}_j\Vert^2/2\sigma^2), \sigma>0$
S 型函数 (sigmoid function)	$k(\boldsymbol{x},\boldsymbol{x}')=\tanh\{\kappa(\boldsymbol{x}\cdot\boldsymbol{x}')+\theta\}$ κ 与 θ 为任意常数

借由选择不同的核函数,即可得到不同的 SVM 非线性判别模型。与人工神经网络模型相比较,非线性 SVM 所得到的最大边缘的超平面与人工神经网络模型相似,例如采用高斯径向基函数所得到的决策界线与使用径向基函数神经网络(radial basis function neural network)相似。核函数的选择并没有特定的规则或方式,用户可根据数据分析结果选择最适当的核函数。

相较于其他分类算法,SVM 的优点在于只要选择适合的核函数,即可处理高维度的非线性分类问题,并且具有良好的分类正确性,此外,由于其最佳的超平面是由少数的支持向量所构成,分析结果也具有较佳的稳健性(robustness)。

10.3 R 语言与随机森林集成学习模型

本节使用皮马族印第安人糖尿病检测数据(Ripley, 1996;Smith *et al.*, 1988)说明如何通过 R 语言构建随机森林集成学习模型进行分类与评估变量重要性。

10.3.1 利用随机森林进行分类

在调用内建于扩充套件 **MASS**(Venables & Ripley, 2002)的数据集后,利用扩充套件 **randomForest**(Liaw & Wiener, 2014)构建随机森林模型。随机森林有两个最主要的参数:ntree 与 mtry,前者设定要产生多少棵决策树作整合(预设为 500 棵),后者则是设定每个决策树分支要使用几个属性(分类默认值为$\sqrt{p}$,p 为数据中的属性数)。以下程序为使用默认值建立模型,并计算测试数据正确率为 0.774。由于随机森林算法具随机性,若程序执行结果可能略有不同。

```
library(MASS)
```

```
library(randomForest)
set.seed(1111) #设定随机数种子
data("Pima.tr")
data("Pima.te")
rf.model<-randomForest(type~ ., data=Pima.tr)
pre.te<-predict(rf.model, Pima.te)
confusion_matrix=table(Pima.te$type, pre.te)
confusion_matrix
test_accuracy=sum(diag(confusion_matrix))/sum(confusion_matrix)#计算正确率
test_accuracy
```

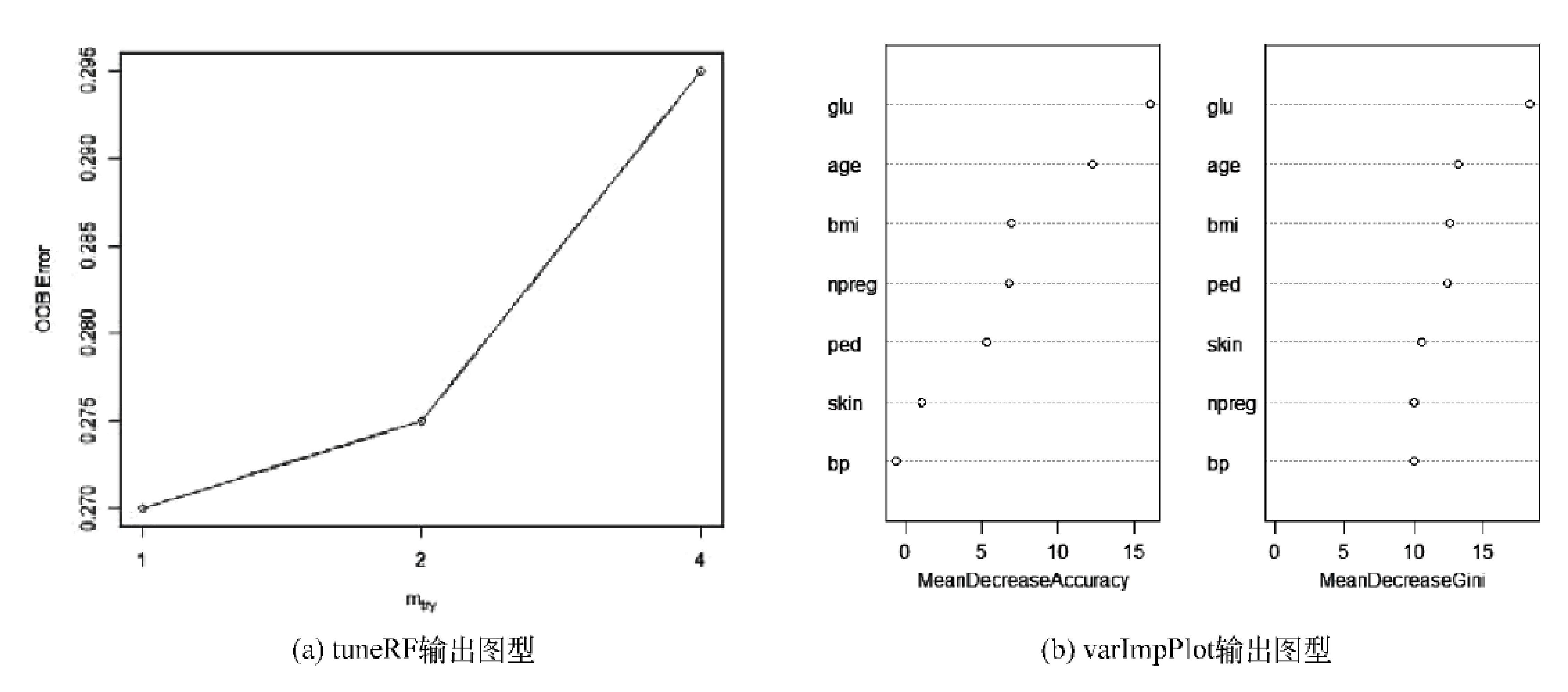
此外，通过 **tuneRF** 函数提供在指定 ntree 设定下进行 mtry 参数微调。以下程序为 ntree＝500 下进行 mtry 参数微调，输出图型如图 10.8(a)所示，以 mtry＝1 为最佳结果。

```
rftune<-tuneRF(y=Pima.tr$type, x=Pima.tr[,1:7], ntreeTry=500)
rf.model<-randomForest(type~ ., data=Pima.tr, ntree=500,
    mtry=rftune[which.min(rftune[,2]),1])
```

10.3.2　利用随机森林评估变量重要性

随机森林算法同时提供评估变量重要性功能，只要在 **randomForest** 函数中设定自变量 importance 为 True 并使用 **varImpPlot** 函数画图，如图 10.8(b)所示。随机森林会用两种指标来排序变量重要性，包含从分类结果来看的正确率下降指标(MeanDecreaseAccuracy)以及从分支不纯度降低指标(MeanDecreaseGini)，前三名重要变量分别为葡萄糖浓度(glu)、年龄(age)、身体质量指数(bmi)。

```
rf.model<-randomForest(type~ ., data=Pima.tr, ntree=500,
    mtry=rftune[which.min(rftune[,2]),1], importance=T)
varImpPlot(rf.model)
```



(a) tuneRF输出图型　　(b) varImpPlot输出图型

图 10.8　随机森林算法输出图型

10.4 结论

随着数据的复杂度越来越高,具有高度学习能力的模型的需求越来越大,集成学习算法与支持向量机可解决传统分类模型准确度不佳的困扰,已被广泛应用在不同的问题上。然而,除了高度的准确度外,与传统决策树分析方法相比较,集成学习算法模型结果的解释度上仍有改善的空间,因此,如何提高可视化的扩展是未来集成学习与支持向量机方法在应用上的重要挑战。

问题与讨论

1. 请比较支持向量机与人工神经网络在二元分类问题上有何异同?

2. 请分析 Bagging 与 Boosting 方法的优缺点。

3. 请根据 20 位受检者的基本资料回答下列各问题。假设有兴趣的目标变量为受检者是否驼背。

(1) 请利用 Bagging 学习算法计算一分类模型(假设最大模型个数 K 为 5)。

(2) 请利用 AdaBoost 建立最后的分类模型(假设学习循环次数为 3)。

(3) 请利用 SVM 建立一分类模型。

(4) 试比较 Bagging、AdaBoost、SVM 三种分类模型的结果与第 4 章利用决策树分析的结果的差异。

心血管疾病数据表

编号	驼背	年龄(>50 岁)	身高(>175cm)	性别	编号	驼背	年龄(>50 岁)	身高(>175cm)	性别
1	是	是	是	男	11	否	否	否	男
2	否	否	是	男	12	否	是	否	女
3	否	是	否	女	13	否	是	否	女
4	否	否	否	女	14	否	否	否	女
5	否	是	否	男	15	否	否	否	男
6	是	是	否	女	16	是	是	是	男
7	否	否	否	男	17	是	是	否	男
8	否	否	否	女	18	否	否	否	男
9	是	否	是	男	19	否	是	否	女
10	否	否	否	女	20	否	是	否	女

第 3 篇

数据挖掘进阶运用

第 11 章

商 业 智 能

11.1 商业智能概述

商业智能(**business intelligence, BI**)是将大量数据转换为具商业价值的信息,以协助企业进行预测、追踪、分析与管理商业行为的工具,使企业能够做较好的决策,本书一开始提到的"尿布与啤酒"就是著名的例子。信息的快速累积与流通带来了更急切的竞争压力,因此,在决策制定过程中,借助大数据分析和商业智能的能力,决定了企业因应商业环境变动的竞争优势。从大数据分析的角度来看,商业智能可说是一种针对商业需求,取得高质量以及有意义的数据挖掘与信息处理机制,以支持商业决策创造企业利益和价值的方法。因此,商业智能关注的是如何整合以及组织数据,并且提供容易使用且可以拿来分享的数据资源,以帮助决策者拟定假设与分析信息、产出结论来减少营运成本,并且加强系统处理的延展性,促使更好的决策产生。

商业智能与数据挖掘、决策的关联性如图 11.1 所示,三者也可视为数据、信息与知识的层级关系。底层乃是企业中各种数据库,整合搜集企业分散于不同系统的数据,例如企业资源计划(enterprise resource planning, ERP)系统、供应链管理(supply chain management,

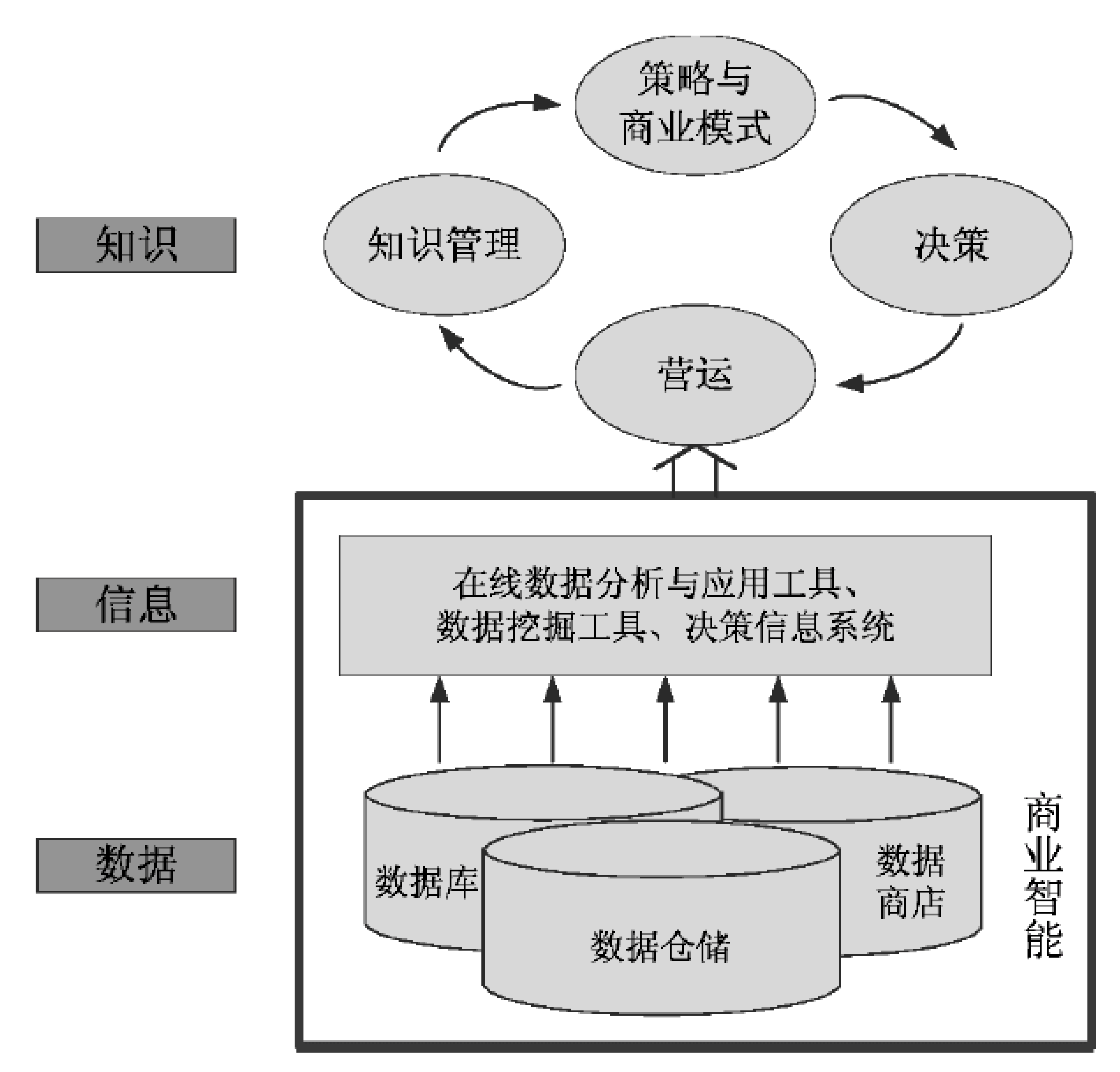


图 11.1　商业智能、数据挖掘与决策

SCM)系统、客户关系管理(customer relationship management, CRM)系统,并利用在线实时分析处理(on line analytical processing, OLAP)与数据挖掘技术将数据转为有价值的信息,辅以报表与查询的功能让对的信息在对的时间传送给对的人。商业智能的范畴一般包含底层的数据搜集到中层的信息产生与传递,如图11.1右下方粗线框的范围。其中,OLAP与数据挖掘在数据转为信息的过程也有所区别,OLAP强调汇整,也就是以不同维度的观点将数据汇整成企业的绩效指标(performance index),让企业管理者借由关键的绩效指标纵观企业整体营运成效,再视需要深入了解指针背后的细节数据;知识管理是通过系统化的管理,将数据转为信息与知识的过程,并将知识予以储存与应用,协助提升企业智能化;数据挖掘更强调新样型探索(exploration),期望借由大量的数据分析,发现新的、能提升营运效率与效能的信息,例如寻找有资金需求且有正常偿债能力客户群的特征、特定购买行为与商场摆设的关系,以及产品良率与特定机台组合的关联性等。

获得有效信息后,由数据挖掘发掘出的规则可纳入专家系统知识库中不断累积新知识,或由领域专家解读从OLAP与数据挖掘获得的数据,进一步结合公司内部拥有的领域专业知识、信息,建立起知识体系,即可纳入企业的知识管理系统将知识储存、扩散及应用。最后结合企业决策者本身的经验与能力,灵活应用知识,即成为企业专属的智能。而企业整体知识的提升将促使各阶层管理者发觉新的决策问题或决策方法,因而产生新的OLAP或数据挖掘主题。决策方法经过验证确认为有效方法后,亦可成为企业知识的另一个来源。

商业智能工具可分为三种类型:①数据汇总软件,主要功能为查询、报表与分析;②数据挖掘和大数据分析工具;③数据市集(data mart)与数据仓储(data warehouse)软件。其中,数据市集与数据仓储已经有多种套装商用软件,因此常被视为另一块独立的领域。而数据汇总软件所具备的“查询、报表与分析”功能则为商业智能中代表性的应用领域,与传统的管理信息系统(management information system, MIS)的差异在于数据汇总软件特别强调多维度分析(multi-dimensional analysis)以及可视化(visualization)的呈现技术。

多维度分析指的是使用者可以依照分析的需求和目标,使用各种不同维度的观点来动态地汇总与呈现数据。相较之下,传统报表内容与更新频率经过信息人员开发完成之后,就不具有变更的弹性,当管理阶层希望从其他角度来分析同一组数据的话,就需要信息人员另外开发一种报表,相对缺乏效率。若能借由数据仓储将数据经由妥善的安排,组织成用户容易理解的存放方法,使用者即可自己选择要分析的数据范围并设计报表内容,迅速取得所需的信息。

可视化则是考虑到人类对于图形和颜色的解读能力,比起对大量数字的解读能力还要来得高。因此,商业智能系统用类似于仪表板(dashboard)的图形化接口在同一个画面中放入数个关键绩效指标(key performance index, KPI),然后以图形来代表数据的差异,并采用颜色管理的方法对每个KPI分别使用不同颜色代表显示范围,例如,以绿色代表绩效良好,黄色代表绩效中等,红色代表绩效不佳,让管理者可以随时掌握企业绩效的全貌。另一方面,可视化功能还包括可以直接点选图形,以进行向下分析(drill-down)来取得更细节的数据,提供管理者掌握现况的信息。

从信息系统的观点来看,内建商业智能和大数据分析模式的**商业智能系统(business intelligence system)**通过信息技术统整散落在不同平台的数据,能结合数据与分析工具,优化决策所需的数据存取与分析,包含基础建设、工具与应用,将复杂且具有竞争力的商业信

息和决策建议呈现给决策者(Watson & Wixom, 2007)。因此企业可以根据管理指标或KPI来汇总数据,并且转变成有用的商业信息以提供决策者进行在线分析处理等数据分析分法,以回答商业问题、预测趋势以及辅助商业决策的系统。例如,许多企业建置"战略室"(war room)以整合相关商业情报和信息;预测企业与产品的走向、费用、资产或是年收入等信息;从累计的数据中建立与分析消费者的信息;预测产业供应链所需的资产并进一步分析出营运过程中可能的风险;借着分析、综合营运、实时互动以对企业的经营绩效进行评估,进而发现潜在的问题或机会;运用大量的数据并且根据区域、单位、产品树状结构等多维度的数据来支持企业决策。

商业智能强调的是提供分析性的营运信息以及简单且多维度的数据查询,以高度可视化的方式呈现信息等特色,因此高阶主管信息系统(executive information system, EIS)亦是商业智能的应用之一,而商业智能和数据挖掘工具就成为提供 EIS 内涵的分析工具,其目的都是要帮助决策者获得足够的信息来架构以及解决决策问题。换言之,单靠功能强大的信息科技并没有办法完全发挥其效用,必须回到分析的本质与目的,针对需求来设计信息的内容,才不会仅是漫无目的地进行数据捞取。以下将探讨商业智能如何应用于交通信息预测、人力资源、机票价格预测与产品需求预测领域的具体个案。

11.2 应用实例——交通信息预测

INRIX 是一家交通路况信息整合公司(http://www.inrix.com),通过与货车、出租车等业者合作,利用智能手机的 App,从 GPS 装置中,将车辆所在位置及移动速率等信息,以匿名方式回传至 INRIX 信息中心,掌握实时的路况与行车信息。

为了提高系统判断交通路线与交通时间的准确率,INRIX 利用大数据来建立各城市的交通流量模型,并且将影响交通的相关因素,例如年度节庆活动、各地气候数据、学校行事历、重要体育赛事等纳入系统之中,使系统可以根据不同的天候状况、特殊节庆或活动的有无,做出更准确的交通信息预测和判断。

此外,INRIX 根据公司搜集的交通信息数据,应用大数据分析与数据挖掘工具,展现出多项创新的商业智能应用与营运模式。例如,在"人潮等于钱潮"的概念思维下,INRIX 分析各大购物商场附近路段的交通拥塞情形,以估计各商场的销售业绩,借由将相关信息贩卖给投资公司,协助投资公司抢在各大购物商场的销售季报或营运年报出炉之前,进行股票买进卖出的投资决策,协助投资公司掌握先机并最大化获利。

另一方面,INRIX 亦将交通路况信息与房地产购买信息两相结合,通过在线地图信息系统,用户可以在搜索并点选有兴趣的房屋地点时,同时看到预计购买的房屋地点与上班地点的实际交通预估时间(actual drive time),提供使用者作为购买房地产的参考信息。

11.3 个案研究——人力资源数据挖掘

11.3.1 案例说明

"人"是企业最重要的资产,人力资源管理和人力资本提升影响组织发展和企业经营绩

效。因此，许多主管都将人力资源管理视为攸关企业生存的决策。其中，“招募与遴选”(recruitment and selection)为最关键的项目。选择对的人才不仅能有效地提升绩效，同时也能促进企业成长与创新。传统遴选方法包括申请表、面谈、智力测验、情境测验或是凭借企业主管个人喜好或经验作为录用准则。然而，随着科技进步、全球化的竞争与组织快速重整与再造，传统人力资源所使用的工作分析与遴选程序显然已不敷使用(Lievens *et al.*，2002)，因此，发展有效的人才遴选方法与规则，以帮助企业管理者找到适才适所的人就成为企业主管重要的课题之一。

本案例(Chien & Chen，2007)应用第8章的粗糙集理论以探索与分析人力相关数据和工作绩效及工作年资，以某半导体制造厂之人才评选机制与数据库作为实证对象，建立人才遴选规则以有效找到与公司文化、工作性质相匹配的优秀人才，具体验证数据挖掘和商业智能在提升人力资本的价值。

11.3.2 分析过程

本案例根据人力资源数据库中找出工作申请者的背景数据与过去工作行为及经验，以粗糙集理论为分析方法，提取人才遴选规则，研究架构如图11.2所示，包含问题定义与数据准备、粗糙集理论分析、规则验证以及所撷取的知识推论。

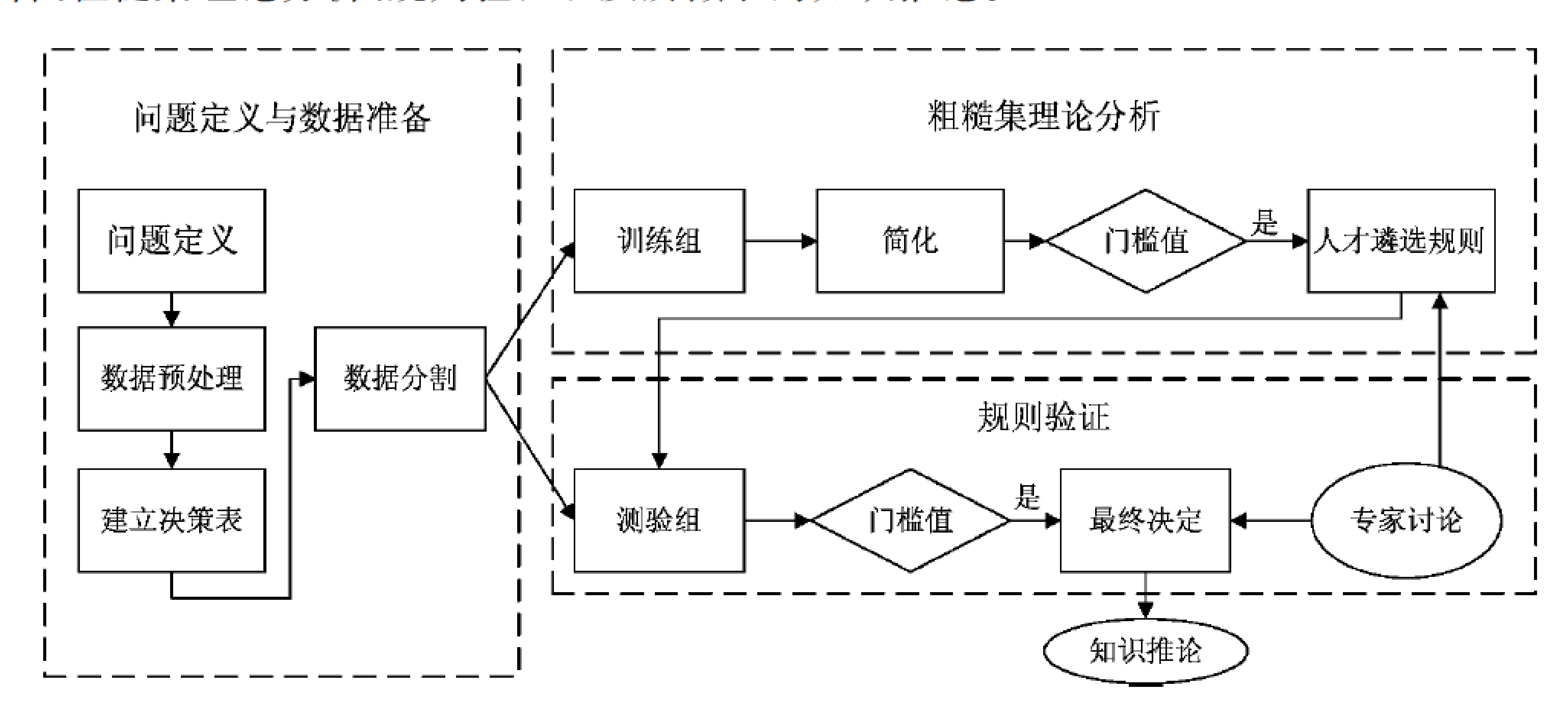


图11.2 粗糙集理论应用于人才遴选的研究架构

1. 问题定义

优秀人才的招募与留任是半导体制造公司维持竞争力的要素之一，随着全球化与技术快速演进，跨部门与跨领域的工作形态崭露头角，工作形态亦因而转换为多样化，遴选具备潜力的员工变得难以仅靠传统的性向测验、工作领域测验、参考过去工作经验以及面试等方法(Chen & Chien，2011；Chien & Chen，2007，2008；简祯富等，2005)。如何吸引并维系公司所需要的人才实属人力资源部门重要的核心工作。决策者首先必须对于公司未来愿景与价值有深刻了解，接续则必须借由列举出该申请人的工作表现与预定目标的差距，以完整了解人事评选机制与员工表现评估结果，并可借由领域专家的协助来定义问题的各个元素与分析方向。

为了找出高潜力的优秀人才，本案例以粗糙集理论为基础，借由人力资源数据库中所记

载的员工过去工作实际表现与目标等比较依据，设计出一套评选机制以选择适任于各种不同工作形态的人才。其输入属性包括年龄、教育程度、工作经验、工作职务以及申请该份工作所使用的管道等。基于保护公司的人事机密，本案例亦将数据与属性给予编码以及筛选后方得建立模式。

由于人力资源数据属于相当机密且隐私的数据群组，数据的储存必须经过设计而存放在不同数据库。因此在数据预处理过程包含确认数据的分布形态、是否有离群值、消除不一致数据、缩减数据维度，并将数据转换成相同格式等步骤，以方便后续的数据分析。完成数据预处理后，则采用随机的方式将数据分成训练组与测验组。前者用于建立模式与规则提取使用，后者则使用于检定所提出架构的效度。

案例公司当时的总员工人数已达 18 570 人，包含 1882 位管理人员、6715 位领域专家、750 位助理工程师及秘书人员、9223 位技师。员工平均年龄为 30.6 岁；有将近一半的员工(46.5%)有大学以上学历；平均服务年资大约为 5 年，并有极高的移转率往其他高科技产业公司就职。在本案例公司的公司规模扩充之际，人才需要与日俱增，纵使每年尚可从各大专院校应届毕业生招募人才，仍须给予新人长期训练，使得公司需付出庞大的人事成本与时间。因此，如何招募到较为适任的新进员工是人事主管重要的任务。

2. 数据准备

本案例先针对工程部门中的五种工作职务，发展人才遴选的机制和规则。所采用的历史数据为 2001 年至 2004 年间此五种工作职务所招募的 3825 位新进员工的人事数据为分析对象，以其过去四年间的工作表现与离职率作为未来招募新进员工的条件取舍。其中，假设所招募的新进员工于一年内即办理离职，则视为人力资源的招募程序不当。因此，须密切分析此类员工的背景、是否适任于该工作性质与其他离职原因，以厘清招募程序是否有误。接着，分析造成离职的因素，以改进人才培训与留任的计划。

本案例所采用的目标变量有工作表现(job performance)、留任(retention)、离职原因(turnover reasons)，分别解释如下。

(1) **工作表现**：本案例公司已建立一绩效评量系统来针对员工当年度表现给予评分，提供管理者与员工了解其过去表现并设定未来年度的绩效表现。根据绩效评分结果，可将员工分为三类，分别为杰出的(outstanding，为前 10%)、成功的(successful，占中间 85%)与尚待改善的(improved needed，为后 5%)。由于被评分为成功的员工占多数，为了避免随机抽样导致分析结果的不正确，因而采用调整后的比例如图 11.3(a)所示，以分层抽样的方式取得 360 位员工的相关数据。

(2) **留任**：由于新进员工的培训成本昂贵，因此新进员工的留职率分析相当重要。可分为两方面探讨：①员工于三个月内辞职属于招募管道的失职；②员工于满三个月至一年内辞职，则归纳于管理阶层与员工培训的失误。然而，若有员工于一年内有意辞职但未辞职则不包含在审视样本中。于是，另外搜集 2622 位员工的相关数据来分析留职率。图 11.3(b)显示此 2622 位员工中于三个月内辞职与否的人数比率，图 11.3(c)则为此 2622 位员工中于一年内辞职与否的人数比率。比较图(b)与图(c)，此样本集中有 15%的员工任职超过三个月但在一年内辞职。

(3) **离职原因**：员工于离职前需与该部门主管以及人力资源部门管理者进行一对一面谈，该离职员工会从 32 种离职原因中选择 3 种可能的原因，然后由其部门主管判断该员工

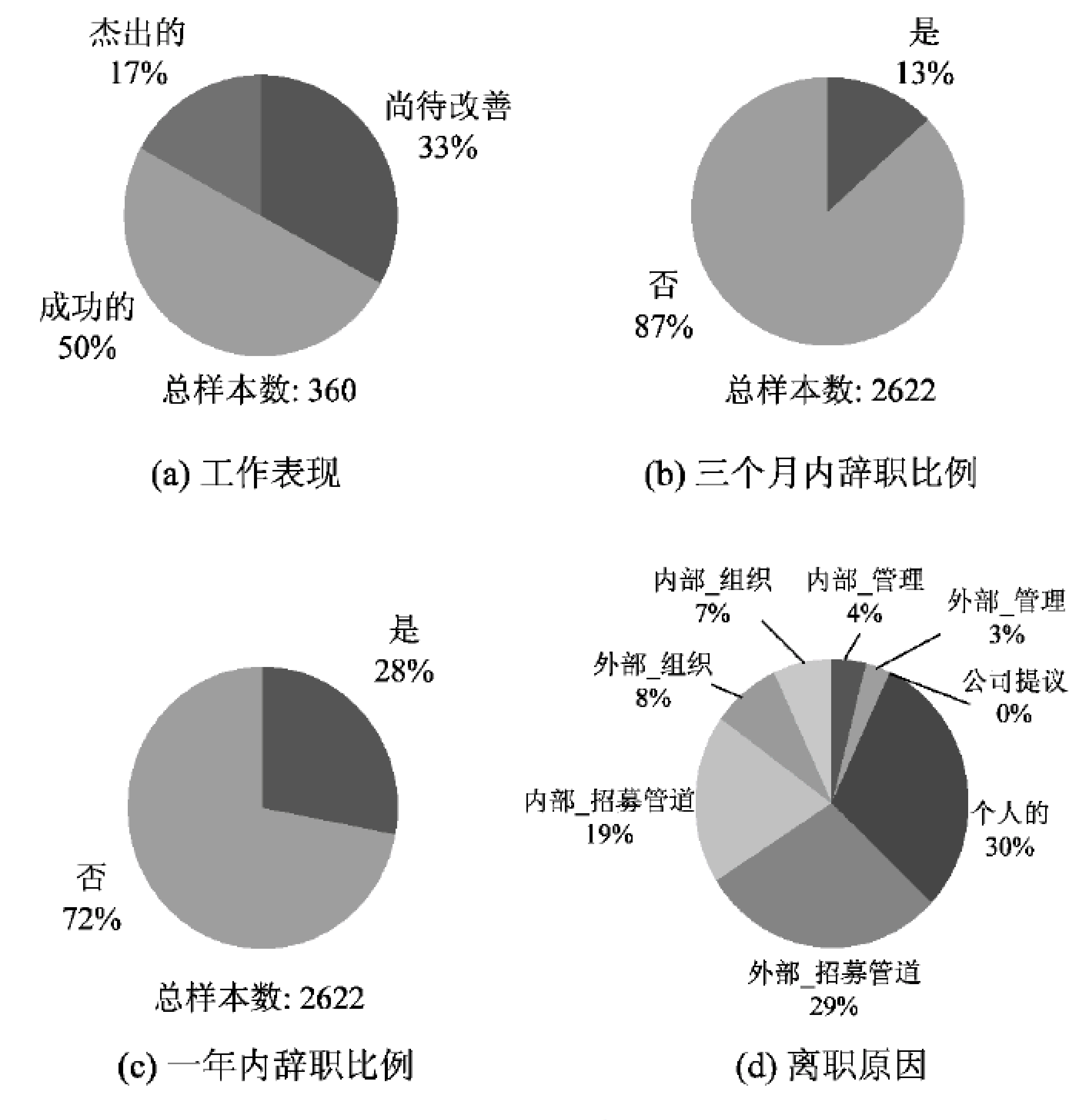


图 11.3　根据不同决策属性的抽样比例

的离职主因，最后则由人力资源部门主管择定该职员的确定离职原因。与领域专家讨论后，此 32 种原因可再归纳为 8 种范畴，如图 11.3(d)所示。

在清楚了解目标变量后，需选择输入变量作为建立模型的基础。一开始，本案例在与领域专家讨论与数据搜集后，找出 9 个输入变量，分别为年龄、性别、婚姻状况、前一份工作年资、教育程度、主修科目范畴、毕业学校、学校排名、招募管道等，定义如表 11.1。然而，基于年龄、性别、婚姻状况等变数会牵涉到性别歧视等议题，故予以删除，最后剩下六个输入变量。

表 11.1　输入变量定义

属性类别	定　　义	变量值定义
年龄	针对受雇者的年龄，分为四个区间类别	1=25 岁或以下 2=26 岁到 30 岁 3=30 岁到 35 岁 4=35 岁以上
性别	受雇者的性别	1=女性 2=男性
婚姻状况	受雇者的婚姻状况	1=单身 2=已婚
前一份工作年资	根据受雇者前一份的工作年资，若超过一年以上则标记为 1，反之则标记为 2	1=超过一年以上 2=一年以内

续表

属性类别	定　　义	变量值定义
教育程度	依照受雇者的受教育程度区分为四个类别	1＝高中职以下 2＝四技二专 3＝大学 4＝研究所以上
主修科目范畴	从目前台湾各大院校的 52 个学科类别，挑出受雇者最多的主修科目，共 11 个科目类别，未被包含的则列入其他项	1＝机械工程 2＝电子工程 3＝化学工程 4＝化学 5＝材料工程 6＝物理学 7＝工程管理 8＝计算机科学 9＝土木工程 10＝环境工程 11＝企业管理 12＝其他
毕业学校	从目前 114 所各大院校，挑出最多受雇者曾就读过的学校，共有 22 所，其中未被包含的则列入其他项	S1～S23，共 23 个类别
学校排名	依据学校排名分为四个类别，前三个类别范畴为依据排名所分的中国台湾学校，第四类别为毕业自海外学校	1＝排名属第一区间 2＝排名属第二区间 3＝排名属第三区间 4＝毕业自海外学校
招募管道	显示招募者当时是通过网络抑或外在招募管道取得公司信息	1＝网络 2＝外在招募管道

3. 粗糙集理论分析

本案例以粗糙集理论提取人才遴选规则，在所有的员工数据中，随机抽取 70％作为训练组样本，剩余 30％则为验证组样本。对应 8.3 节的八个步骤，分析结果如下：

步骤 1：在图 11.3 的(a)与(c)中，可以发现人才留任比例并不平衡。因此依照比例大小而对候选规则有不同的筛选条件：①当产生的规则中包含比例大的样本，则必须要有五个样本以上才能够支持该候选规则，如规则包含“三个月内未辞职”或“一年内未辞职”，则需至少五个样本来支持；②当产生的规则中包含比例小的样本，则仅需两个样本即能够支持这个候选规则，如规则包含“三个月内辞职”或“一年内辞职”，则仅需两个样本来支持。

步骤 2：建立工作表现、一年内未辞职、三个月内未辞职、离职因素等四个决策表。

步骤 3：由于本案例所采用的属性皆为离散型，因此忽略此步骤而直接进入步骤 4。

步骤 4～6：利用训练组样本产生规则，包含 640 条评估工作表现的规则、622 条评估一年内辞职的规则、519 条评估三个月内辞职的规则、2959 条评估离职原因的规则。

步骤 7 和 8：与领域专家讨论所搜集的候选规则集合，剔除不符合实务的规则，则各剩 156 条评估工作表现的规则、135 条评估一年内辞职的规则、91 条评估三个月内辞职的规则、547 条评估离职原因的规则。

表 11.2～表 11.5 分别为以工作表现、一年内辞职、三个月内辞职以及离职原因作为输

出变量所产生的候选规则。

表 11.2 以“工作表现”变量为输出变量所产生候选规则的支持度

候选规则	输入变量							输出变数	支持度
	工作职等	招募管道	工作经验	教育程度	毕业学校	学校排名	主修科目	工作表现	
1	F2	内部	是	—	—	2	—	杰出	2
2	—	内部	是	硕士以上	S3	—	—	杰出	2
3	F2	—	—	—	—	1	化学	一般	5
4	F5	外部	是	—	—	—	—	一般	5
5	F3	外部	—	—	—	2	—	尚待改善	2
6	—	外部	—	学士	—	4	电机工程	尚待改善	2

表 11.3 以“一年内辞职”变量为输出变量所产生候选规则的支持度

候选规则	输入变量							输出变数	支持度
	工作职等	招募管道	工作经验	教育程度	毕业学校	学校排名	主修科目	一年内辞职	
1	F5	—	无	—	—	—	—	否	139
2	F2	外部	—	—	S2	—	—	否	36
3	F4	外部	—	—	—	1	—	否	30
4	F1	外部	是	硕士以上	—	2	—	是	5
5	F3	外部	是	—	S3	—	—	是	3

表 11.4 以“三个月内辞职”变量为输出变量所产生候选规则的支持度

候选规则	输入变量							输出变数	支持度
	工作职等	招募管道	工作经验	教育程度	毕业学校	学校排名	主修科目	一年内辞职	
1	F1	—	—	—	S6	—	—	否	27
2	F1	—	无	—	S5	—	—	否	20
3	—	内部	无	—	—	4	—	否	19
4	F1	—	是	硕士以上	S15	—	—	是	2
5	—	—	无	学士	S2	—	机械工程	是	2

表 11.5 以“离职原因”变量为输出变量所产生候选规则的支持度

候选规则	输入变量									输出变数	支持度
	工作职等	招募管道	工作经验	教育程度	毕业学校	学校排名	主修科目	工作表现	留职情形	离职原因	
1	F3	—	无	—	—	—	材料工程	—	三个月内辞职	个人因素	4
2	—	外部	—	—	—	—	电机工程	一般	一年内辞职	外部_招募管道	4
3	—	内部	是	—	—	—	物理	一般	—	外部_组织	3

续表

候选规则	输入变量									输出变数	支持度
	工作职等	招募管道	工作经验	教育程度	毕业学校	学校排名	主修科目	工作表现	留职情形	离职原因	
4	F1	外部	—	硕士以上	S4	—	—	—	三个月内辞职	外部_招募管道	3
5	F5	—	—	—	—	3	—		一年内辞职	内部_管理	2

4. 规则验证与知识推论

通过与该领域专家讨论，找出最符合实务情况与适切的候选规则，再以置信度与增益值作为评选适切规则的门槛值，并根据上述四个步骤完成此一验证程序。

步骤 1：将两门槛值置信度与增益值，分别设定为 90%与 1，作为遴选规则的标准。

步骤 2：以测试组的样本进行模式验证。通过此一机制的候选规则，则形成“If-Then”的规则形态。

步骤 3：通过与预先设定的门槛值做比较(置信度为 90%、增益值为 1)，选出最终用来遴选人才的规则。每一个测验组的样本为输入值，以验证所有的候选规则。最后再以工作表现、一年内未辞职、三个月内未辞职、离职因素等四种变量为输出变量所建立的规则进行验证，各找出显著 If-Then 规则有 9、31、36、11 条，如表 11.6～表 11.9 所简述。

步骤 4：与领域专家讨论并检查所有通过检验的候选规则，并无发现不适用的规则，因此本案例所发掘的 87 条规则将全数保留。

表 11.6 以“工作表现”变量为输出变量所产生候选规则的信度验证

候选规则	If-Then 规则形式	满足前提条件的样本数	满足前提条件且决策结果的样本数	置信度	增益	规则接受与否
1	若申请者的申请工作职等为 F2、有相关工作经验、从内部招募管道申请、毕业自第二种群组的学校 则可推论此申请者于未来工作表现为优异	1	1	100%	7.27	Yes
2	若申请者的有相关工作经验、从内部招募管道申请、毕业自排名为 S3 学校、具有硕士以上的学历 则可推论此申请者于未来工作表现为优异	1	1	100%	7.27	Yes
3	若申请者的申请工作职等为 F2、有相关工作经验、毕业自第二种群组的学校、主修化学 则可推论此申请者于未来工作表现为一般	6	3	50%	0.99	No
4	若申请者的申请工作职等为 F5、有相关工作经验、从外部招募管道申请 则可推论此申请者于未来工作表现为一般	1	1	100%	1.98	Yes
5	若申请者的申请工作职等为 F3、从外部招募管道申请、毕业自第二种群组的学校 则可推论此申请者于未来工作表现将尚待加强	1	1	100%	2.79	Yes

续表

候选规则	If-Then 规则形式	满足前提条件的样本数	满足前提条件且决策结果的样本数	置信度	增益	规则接受与否
6	若申请者从外部招募管道申请、具有大学学历、毕业自第四种群组的学校 则可推论此申请者于未来工作表现将尚待加强	3	2	67%	1.86	No

表 11.7 以“一年内辞职”变量为输出变量所产生候选规则的信度验证

候选规则	If-Then 规则形式	满足前提条件的样本数	满足前提条件且决策结果的样本数	置信度	增益	规则接受与否
1	若申请者的申请工作职等为 F5、无相关工作经验 则可推论此申请者于一年内将不会辞职	65	60	92%	1.26	Yes
2	若申请者的申请工作职等为 F2、无相关工作经验、从外部招募管道申请、毕业自排名为 S2 学校 则可推论此申请者于一年内将不会辞职	20	17	85%	1.16	No
3	若申请者的申请工作职等为 F4、从外部招募管道申请、毕业自第一种群组的学校 则可推论此申请者于一年内将不会辞职	17	12	71%	0.97	No
4	若申请者的申请工作职等为 F1、从外部招募管道申请、有相关工作经验、具有硕士以上学历、毕业自第二种群组的学校 则可推论此申请者将会于一年内辞职	2	2	100%	3.71	Yes
5	若申请者的申请工作职等为 F3、从外部招募管道申请、有相关工作经验、毕业自排名为 S3 学校。 则可推论此申请者将会于一年内辞职	5	4	80%	2.97	No

表 11.8 以“三个月内辞职”变量为输出变量所产生候选规则的信度验证

候选规则	If-Then 规则形式	满足前提条件的样本数	满足前提条件且决策结果的样本数	置信度	增益	规则接受与否
1	若申请者的申请工作职等为 F1、毕业自排名为 S6 学校 则可推论此申请者于三个月内将不会辞职	10	10	100%	1.15	Yes
2	若申请者的申请工作职等为 F1、无相关工作经验、毕业自排名为 S5 学校 则可推论此申请者于三个月内将不会辞职。	7	7	100%	1.15	Yes
3	若申请者从内部招募管道申请、无相关工作经验、毕业自排名为 S4 的学校 则可推论此申请者于三个月内将不会辞职	7	7	100%	1.15	Yes

续表

候选规则	If-Then 规则形式	满足前提条件的样本数	满足前提条件且决策结果的样本数	置信度	增益	规则接受与否
4	若申请者的申请工作职等为 F1、有相关工作经验、具有硕士以上学位、毕业自排名为 S15 学校 则可推论此申请者将会于三个月内辞职	2	1	50%	3.78	No
5	若申请者无相关工作经验、具有大学学历、毕业自排名为 S2 学校、主修机械工程 则可推论此申请者将会于三个月内辞职	1	1	100%	7.57	Yes

表 11.9　以“辞职原因”变量为输出变量所产生候选规则的信度验证

候选规则	If-Then 规则形式	满足前提条件的样本数	满足前提条件且决策结果的样本数	置信度	增益	规则接受与否
1	若申请者的申请工作职等为 F3、无相关工作经验、主修材料工程、于三个月内辞职 则可推论此申请者的离职原因为基于个人因素	1	1	100%	3.32	Yes
2	若申请者从外部招募管道申请、主修电子工程、工作表现为一般、于一年内辞职 则可推论此申请者的离职原因为基于公司内部压迫导致	3	2	67%	2.19	No
3	若申请者有相关工作经验、主修物理学、工作表现为一般 则可推论此申请者的离职原因为基于公司外部环境吸引	4	1	25%	3.71	No
4	若申请者的申请工作职等为 F1、从外部招募管道申请、具有硕士以上的学位、毕业自排名为 S4 学校、于三个月内辞职 则可推论此申请者的离职原因为基于公司外部环境吸引	1	1	100%	3.28	Yes
5	若申请者的申请工作职等为 F5、毕业自第三种群组的学校、于一年内辞职 则可推论此申请者的离职原因为基于公司内部压迫导致	2	1	50%	14.1	No

11.3.3　案例小结

本案例发现员工表现与背景之间的关联。例如，虽然一般认为，毕业自顶尖大学的学生表现应较为优异，然而，本研究分析发现在设备维修工作职务上表现较佳与留职时间较长的员工，反而是毕业于一般大学的学生。因此，本案例公司已与相关大学科系，建立产学合作研究机制与暑期学生实习等，以吸引“志同道合”的人才。此外，由分析结果亦显示，对于某些工作性质而言，由内部推荐管道所招募的人才表现会相对较佳，公司因此设计激励奖金的机制，鼓励内部员工推荐的招募管道，以提升人才招募的效率与质量。通过此机制，不仅可以在第一时间替公司招募适合该工作性质的人才，也可以挽留人才并提升员工对工作的满足感。再者，此一机制亦提供人力资源部门进行有效的人事管理，如工作内容设计、内部工作轮替、监控以及工作职能训练等。

高科技产业有赖于能适应动态的工作性质的人才来维系公司的市场竞争力。本案例采用粗糙集理论作为建立数据挖掘模型和人才评选分类规则的理论基础，分析现有员工绩效表现与其背景数据，以提取潜在有价值的分类规则，适宜应用于作业阶层或管理阶层的人才遴选。通过应用商业智能的评选机制，可初步筛选不适任的求职者，以节省公司人才训练成本。此外，本书第4章介绍的决策树也常用以发掘潜在的If-Then规则(Chien & Chen, 2008)，并可混合第10章支持向量机算法以提升分析成效(Chen & Chien, 2011)。

11.4 应用实例——机票价格预测

Farecast是一个机票价格预测网站(http://www.farecast.com/)，2002年美国华盛顿大学埃齐奥尼(Oren Etzioni)教授在一次搭机途中与隔壁乘客闲聊时询问对方的机票买了多少钱，一问之下，发现自己虽然比起其他人更早购买机票，但是却买得比别人贵。他开始调查同班飞机其他乘客购买机票的价格，发现于不同时间购买相同舱等的机票价格并不一致，因此，他决定运用大数据分析与数据挖掘技术以建构机票价格波动的预测模型，作为预测机票价格未来涨跌的指引。他设立了一个网站，从各个管道搜集机票价格数据，并由用户的购买经验与累积的历史数据，分析未来一周最佳的机票购买时间，以提供使用者何时可以购买最便宜的机票的相关信息(Darlin, 2006)。

使用者只要输入出发与预计到达的时间地点以及人数，Farecast即可从庞大的数据库中分析归纳出当下各家航空公司的机票价格，并以不同颜色的箭号显示该价格在未来将会上涨或下跌，并给予使用者应该“现在直接购买”或“稍候再购买”的购买建议，以在浮动的机票价格中做出最优惠的购买决策。

Farecast进一步设计一个创新性的机票购买服务与商业模式：“保证价格”(fare guard)。其运作方式是，当Farecast网站预测某一机票价格于未来将维持不变或倾向下跌时，在当下还未决定是否购买机票的顾客，则可先向Farecast购买“保证价格”(约9.95美元)，以保证于未来一周内皆可以“当日最低价”来购买机票，即使未来一周该机票价格上涨，已购买“保证价格”的用户，也能获得保证价格与购买价格之间的差价作为补偿。Farecast机票价格预测网站的服务因此备受欢迎，成为大数据分析与商业智能极佳的应用范例之一。因此，2008年微软出价1.15亿美元收购此公司。

11.5 个案研究——产品需求预测

需求预测是生产决策和规划的基础，准确的需求预测可降低存货成本、维持顾客订单服务水平、协助产能规划及提升资本效益，进而增加公司竞争力。随着半导体进入消费性电子产品时代，更多元的产品可供消费者选择，同时也缩短了产品生命周期和产品替代时间。另一方面，半导体产品的多样性消费受到了各种经济因素的影响，进而增加了市场供需的波动和需求预测的风险，使得半导体产品需求的预测变得更加困难和复杂。

11.5.1 半导体产品需求预测架构

本案例(Chien *et al.*, 2010)以产品生命周期与技术扩散模型为基础，提出半导体产品的需求预测架构，共包含六个阶段：问题定义、数据准备、构建扩散模型、参数估计、预测、解释和验证，如图11.4所示。

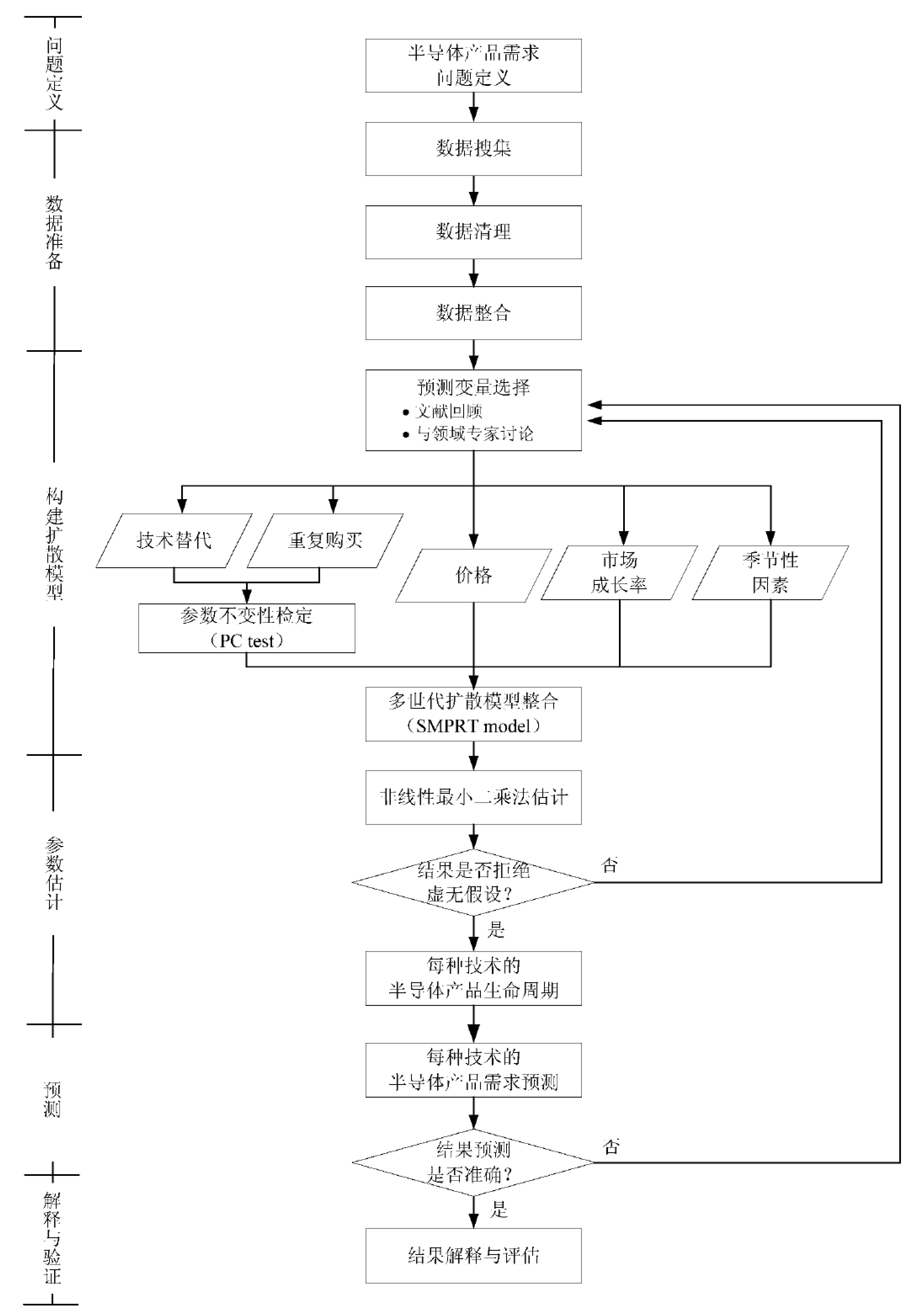


图 11.4　产品需求预测的架构(Chien *et al.*，2010)

本案例中使用的符号定义如下：

$F_i(t)$　产品世代 i 在时间 t 的累积密度函数

$f_i(t)$　产品世代 i 在时间 t 的概率密度函数

$s_i(t)$　产品世代 i 在时间 t 的实际销售量

$X_i(t)$　产品世代 i 在时间 t 的累积的市场效应

$\hat{s}_i(t)$　产品世代 i 在时间 t 的估计销售量

τ_i　产品世代 i 的推出时间，$\tau_i \geqslant 1$

p_i　产品世代 i 的创新系数(innovation coefficients)，代表该世代的创新者比例

q_i　产品世代 i 的模仿系数(imitative coefficients)，代表该世代模仿者比例

m_i　产品世代 i 的市场潜力(incremental market potential)

ρ_i　产品世代 i 的平均重复购买率

M_i　产品世代 i 的总市场潜力 $M_i = m_i \times \rho_i$

α_t　时间 t 的季节性因素

β　价格系数

$pr_i(t)$　产品世代 i 在时间 t 的价格

g_t　时间 t 的市场成长率

n　世代的数量

l　期间的数量

1. 数据准备

提升数据质量是构建预测模型的重要关键，对于需求事务数据库所累积的大量历史数据往往需确保是否存在遗漏值等数据格式不一致的情况。数据清理的过程包括删除或填补遗失数据，数据整合则是将产品价格、半导体产业的市场成长率和产品需求等来源不同的数据汇整成为一份分析数据。

2. 建构需求预测模型

巴斯(Bass，1969)提出技术扩散模型作为描述技术创新扩散的过程，将产品采用者分为两种：一种是受到大众媒体传播影响的先驱采用者，又称为创新者(innovator)，另一种是受到口耳相传影响的追随采用者，又称为模仿者(imitator)。Bass 模型是依据产品生命周期预测新产品首次被购买的销售量，并假设采用者在创新的过程只能购买一次，没有重复购买的情形。

本案例以 Norton & Bass(1987)模型为基础建立一个多世代扩散模型。基于假设检定以判断模型参数在不同世代间是否会有显著差异，并加入影响半导体产品需求的影响因子作为模式建构，例如产品价格和半导体市场成长率。

本案例提出多世代扩散模型考虑下列五项因素，称为 **SMPRT 多世代技术扩散模型**。

(1) **季节性因素(season factors)**：消费性产品的销售往往容易受不同季节影响，例如对计算机制造业和通信产业来说，圣诞节前后计算机的销售量一般多会显著增加。因此，必须考虑季节性因素以免高估或低估需求。

(2) **市场成长率(market growth rates)**：市场成长率是用来描述市场结构和经济环境。成长率会影响顾客行为的变化，而需求和市场成长率呈现正比关系。

(3) **价格(price)**：在多世代产品的情况下，价格是影响客户购买的重要因素。定价策略不仅要考虑生命周期的阶段，也要考虑新一代产品的替代。价格会影响产品的需求，价格上升则需求下降，反之亦然。新一代产品推出时的价格会影响顾客购买的意愿，顾客的行为会根据未来价格是否如预期涨跌而改变。以半导体产品来说，在初期时价格会迅速下降，并在后期呈现稳定状态。巴斯等(Bass *et al.*，1994)将市场效应纳入单一世代的 Bass 模型中，其市场效应包含价格和广告支出，本案例将此模型进一步扩大为多世代模型。

(4) **重复购买(repeat purchases)**：巴斯等(Bass & Bass，2001)主要是修改诺顿和巴斯(Norton & Bass，1987)提出的多世代扩散模型，两个模型中均考虑重复购买的因素，且分为两部分，以先驱采用者和追随采用者为代表。

(5) **技术替代效应(technological substitution effect)**：半导体产业快速发展，不断被引进并推向市场的新技术不仅逐步取代旧技术，同时也扩大市场潜力。新的半导体产品在本质上是技术创新，必须同时考虑扩散和替代，大多数的预测方法只着重在新技术，却忽略了旧一代产品可能会和新产品竞争，因此单代扩散模型或其他预测方法并不适用于半导体产品，而需采用多世代扩散模型。

首先，Norton & Bass(Norton & Bass，1987)模型以及 Islam & Meade 模型(Islam & Meade，1997)如式(11.1)所示：

$$s_i(t)=f_i(t)[M_i+f_{i-1}(t)[M_{i-1}+f_{i-2}(t)[M_{i-2}+\cdots+f_2(t)[M_2+f_1(t)M_1]\cdots]]] \cdot [1-f_{i+1}(t)] \tag{11.1}$$

其中，$f_i(t)=F_i(t)-F_i(t-1)$，

$$F_i(t)=\begin{cases}\dfrac{1-e^{-(p_i+q_i)(t-\tau_i+1)}}{(q_i/p_i)e^{-(p_i+q_i)(t-\tau_i+1)}+1}, & t\geq\tau_i\\ 0, & t<\tau_i\end{cases}$$

Norton & Bass(1987)与 Islam & Meade(1997)皆为多世代扩散模型，其差别在于参数在跨世代中是否改变。在 Norton & Bass(1987)的模型中，参数在跨世代中是固定的，表示为

$$F_i(t)=\frac{1-e^{-(p_i+q_i)(t-\tau_i+1)}}{(q_i/p_i)e^{-(p_i+q_i)(t-\tau_i+1)}+1}=\frac{1-e^{-(p+q)(t-\tau_i+1)}}{(q/p)e^{-(p+q)(t-\tau_i+1)}+1} \tag{11.2}$$

而 Islam & Meade(1997)的模型中，参数在跨世代中是随着不同期数而变动，表示为

$$F_i(t)=\frac{1-e^{-(p_i+q_i)(t-\tau_i+1)}}{(q_i/p_i)e^{-(p_i+q_i)(t-\tau_i+1)}+1} \tag{11.3}$$

其中，$p_1=p_1, p_i=p_{i-1}+\Delta p_i, i=2,3,\cdots$

用 SMPRT 模型来测试跨代间参数是否改变的步骤如下(Chien *et al.*，2010)：

(1) 定义 Norton & Bass 模型为受限制模型。

定义 Islam & Meade 模型为非受限制模型。

(2) 假设：

$$H_0: \Delta p_i=0 \text{ 且 } \Delta q_i=0\ \forall i$$
$$H_1: \Delta p_i^2+\Delta q_i^2>0$$

(3) 检定：

在 H_0 是正确的假设下，可推导检定统计量：

$W=2$（log max likelihood（未被限制的模型）－log max likelihood（被限制的模型））为近似自由度为 v 的卡方分布，v 为限制与未限制模式参数自由度的差异。

此假设可以借由饱合信息最大似然估计(full information maximum likelihood，FIML)来进行检定，可以利用过去的历史需求数据来检验统计，由检测的结果便可以决定是否要拒绝虚无假设 H_0。

$W_0=2$(log likelihood(未被限制的模型)－log likelihood(被限制的模型))。

(4) 检验规则，此步骤决定是否拒绝虚无假设。

① 如果 $W_0>\chi_\alpha^2(v)$，则拒绝虚无假设 H_0。显示有足够的证据当多个世代交替时，p_i 和 q_i 的值会改变。

② 如果 $W_0<\chi_\alpha^2(v)$，则接受虚无假设 H_0。表示没有足够的证据去证明当多个世代交替时，p_i 和 q_i 的值会改变。

(5) 结果：根据上述的假设检定，可得到 p_i 和 q_i 在不同世代间是否会改变。

此外，本案例研究所发展的 SMPRT 模型也纳入了价格的因素，Norton & Bass 模型中的 $F(t)$取代了巴斯等(Bass *et al.*，1994)提出的 $F(t)$，修改如式(11.4)与式(11.5)所示：

$$F_i(t)=\frac{1-\mathrm{e}^{-(X_i(t)-X_i(0))(p_i+q_i)}}{(q/p)\mathrm{e}^{-(X_i(t)-X_i(0))(p_i+q_i)}+1} \tag{11.4}$$

$$X_i(t)=t+\ln\left(\frac{pr_i(t)}{pr_i(0)}\right)\beta \tag{11.5}$$

市场成长率和季节性因素是半导体产品的需求预测重要的因素。古典的时间序列分解可以用加法和乘法模型分析时间序列数据组成，其分解方法将变量分为四个部分，分别为：长期趋势、季节变化、周期性波动和不规则波动。本研究的 SMPRT 模型采用乘法模型来表达季节性变化和市场成长率的影响如下：

(1) 修改后的产品需求：$\hat{s}_i(t)\times\alpha_t\times\exp(g_t)$

(2) 利用非线性最小二乘法估计参数并修订方程式如式(11.6)与式(11.7)所示：

$$\text{原始：Min}\sum_{i=1}^{n}\sum_{t=1}^{l}[s_i(t)-\hat{s}_i(t)]^2 \tag{11.6}$$

$$\text{修正后：Min}\sum_{i=1}^{n}\sum_{t=1}^{l}[s_i(t)-\hat{s}_i(t)\times\alpha_t\times\exp(g_t)]^2 \tag{11.7}$$

$$\text{而 }\alpha_1=1,\alpha_t=\alpha_{t-4},\alpha_t\geqslant 0,\quad g_t\text{：常数}$$

将 SMPRT 模型整合成如下的方程式，如式(11.8)所示：

$$\text{Min}\sum_{i=1}^{n}\sum_{t=1}^{l}[s_i(t)-\hat{s}_i(t)\cdot\alpha(t)\cdot\exp(g(t))]^2 \tag{11.8}$$

限制于

$$s_i(t)=f_i(t)[M_i+f_{i-1}(t)[M_{i-1}+f_{i-2}(t)[M_{i-2}+\cdots+f_2(t)[M_2+f_1(t)M_1]\cdots]]]\cdot[1-f_{i+1}(t)]$$

$$f_i(t)=F_i(t)-F_i(t-1)$$

$$F_i(t)=\begin{cases}(1-\mathrm{e}^{-(X_i(t)-X_i(0))(p_i+q_i)})/[(q_i/p_i)\mathrm{e}^{-(X_i(t)-X_i(0))(p_i+q_i)}+1], & t\geqslant\tau_i\\ 0, & t<\tau_i\end{cases}$$

$$X_i(t)=(t-\tau_i+1)+\ln(pr_i(t)/pr_i(0))\beta$$

$$\alpha_1 = 1,\quad \alpha_t = \alpha_{t-4},\quad \alpha_t \geqslant 0$$
g_t：常数
$$0 < p_i < 1$$
$$0 < q_i < 1$$
$$M_i > 0$$
$$\forall\, i = 1,2,\cdots,n$$

3. 参数估计

SMPRT 模型需根据历史需求数据来估计参数。本案例以非线性最小二乘法来估计参数，并分为两阶段。首先，分别估计 Norton & Bass 模型以及 Islam & Meade 模型里的参数；并利用 FIML 检定参数是否不变；其次，再估计在 SMPRT 模型中的参数，包含创新系数(p_i)、模仿系数(q_i)、市场总潜力(M_i)、季节性因素(α_t)、价格有效性(β)。

4. 结果解释与评估

本研究将数据分为训练组和测试组，训练组用来估计 SMPRT 模型的参数和之后的需求预测，测试组用来验证需求预测的结果以及和实际需求做比较。

预测误差是评价预测绩效的标准，可以用几个指标来衡量，例如，均方误差(mean square error，MSE)、绝对平均误差(mean absolute error，MAE)和绝对平均百分比误差(mean absolute percentage error，MAPE)等。刘易斯(Lewis，1982)建议用 MAPE 来评估绩效的标准，如表 11.10 所示，较小的 MAPE 值表示在未来需求预设的准确性越高。MAPE 的公式计算可见 10.5.2 节的说明。

表 11.10 MAPE 绩效建议参考表

预测绩效	很好	可接受	尚可	不精确
MAPE/%	<10	10～20	20～50	>50

11.5.2 分析过程

本案例以半导体产品需求预估为实证，以作为半导体厂产能规划的决策依据，协助商业决策提升公司获利。半导体公司通常采用订货型生产方式(make-to-order)，其产品的数量会影响到生产的量和分配，若产量大于需求会导致机器空转和资源浪费；反之，若产量小于需求会导致订单流失和损害商誉。由于新制程技术缺乏需求预测的信息，对于新技术产品的需求预测变得更加复杂且困难。

过去产品需求的事务数据是以季为单位存在数据库中，首先要先进行数据准备，搜集、清理、分割及整合数据。本案例以半导体公司的一主导产品 X 做验证，并使用转换后的数据。此半导体产品 X 之数据总共有 36 个季度，以 4 种不同技术世代制造的产品 X。

在 SMPRT 模型中需要的数据报括产品价格、半导体产业成长率和产品需求量且须先删除不需要的数据如表 11.11 所示。计算产品 X 在相同的技术与季度下的总数量，如表 11.12 所示。

表 11.11 产品 X 的需求整合

顾 客	产 品	技 术	200×1Q	200×2Q
1	X	A	50	30
2	X	B	25	30
1	X	C	50	80
2	X	C	70	90
3	X	D	70	105
3	X	B	58	70
1	X	D	80	100

表 11.12 产品 X 的需求整合

产 品	技 术	200×1Q	200×2Q
X	A	50	30
X	B	83	100
X	C	120	170
X	D	150	200

此数据包含了过去 X 产品在不同技术下制造的需求量、单位价格、季度销售成长率、季度销售成长率的财务报告，如图 11.5 所示。

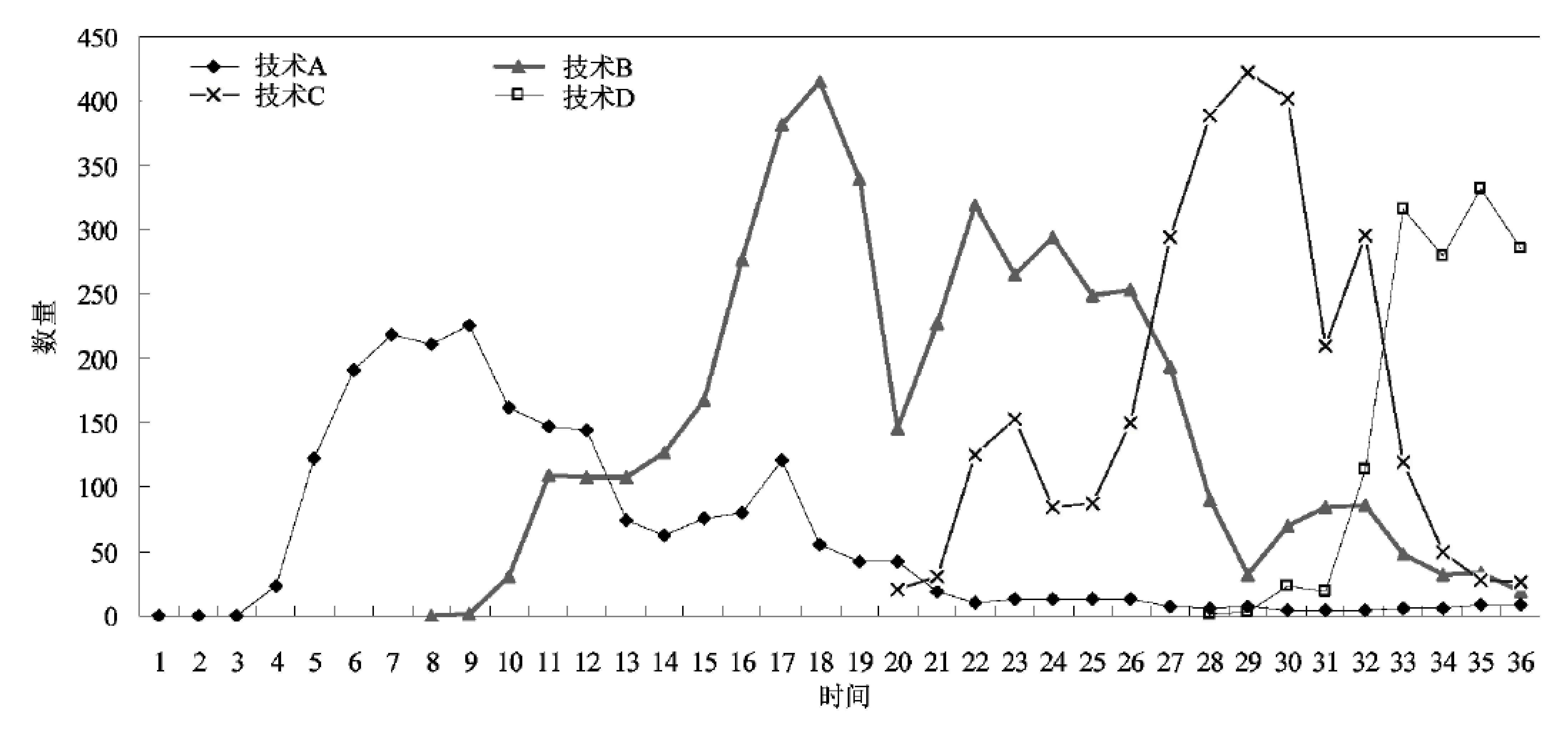


图 11.5 不同技术制造的 X 产品的需求量

1. 建立 SMPRT 模型

步骤 1：检验创新系数和模仿系数在跨世代是否不变。比较 Norton & Bass 以及 Islam & Meade 模型。产品 X 的四个技术世代建立的扩散模型如式(11.9)所示：

$$\left.\begin{aligned}
s_4(t) &= f_4(t)[M_4 + f_3(t)[M_3 + f_2(t)[M_2 + f_1(t)M_1]]] \\
s_3(t) &= f_3(t)[M_3 + f_2(t)[M_2 + M_1 f_1(t)]][1 - f_4(t)] \\
s_2(t) &= f_2(t)[M_2 + M_1 f_1(t)][1 - f_3(t)] \\
s_1(t) &= M_1 f_1(t)[1 - f_2(t)]
\end{aligned}\right\} \tag{11.9}$$

其中，

$$f_i(t) = F_i(t) - F_i(t-1)$$

$$\tau_1 = 1, \tau_2 = 8, \tau_3 = 20, \tau_4 = 28$$

$$i = 1,2,3,4$$

$$F_i(t) = \begin{cases} \dfrac{1 - e^{-(p_i+q_i)(t-\tau_i+1)}}{[(q_i/p_i)e^{-(p_i+q_i)(t-\tau_i+1)} + 1]}, & t \geqslant \tau_i \\ 0, \quad t < \tau_i \end{cases}$$

根据 Norton & Bass 模型，创新系数和模仿系数在每个世代中均不变：

$$p_1 = p_2 = p_3 = p_4$$

根据 Islam & Meade 模型，创新系数和模仿系数不断地改变：

$$p_1 = p_1, \quad p_2 = p_1 + \Delta p_2, \quad p_3 = p_2 + \Delta p_3, \quad p_4 = p_3 + \Delta p_4$$

步骤 2：检验测试参数是否不变。

(1) 定义 Norton & Bass 模型为受限制模型，定义 Islam & Meade 模型为非受限制模型。

(2) 假设：

$$H_0: \Delta p_i = 0 \text{ 且 } \Delta q_i = 0 \ \forall i, \quad i = 2,3,4$$

$$H_1: \Delta p_i^2 + \Delta q_i^2 > 0, \quad i = 2,3,4$$

(3) 检定：

$$W = 2(\text{log max likelihood(非限制模型)} - \text{log max likelihood(被限制模型)}) \sim \chi^2(4)$$

(4) 使用 FIML 函数得到受限制及非受限制模型的对数似然函数(log likelihood)，如表 11.13 所示，结果如下：

$$W_0 = 2[-701.0875 - (-749.3858)] = 96.5966$$

表 11.13　log likelihood 结果

模　　型	Norton & Bass	Islam & Meade
log likelihood	−749.3858	−701.0875

(5) 从卡方分配临界值表得到 $\alpha = 0.05$ 和自由度 $= 4$。

$$\chi^2_{0.05}(4) = 9.488$$

$$\chi^2_{0.05}(4) = 9.488 < W_0 = 96.5966$$

(6) 由于 $\chi^2_{0.05}(4) < W_0$，所以拒绝虚无假设 H_0，因此可得知当世代交替时，p_i 和 q_i 有显著改变。

步骤 3：SMPRT 多世代模型。若以 X 产品在第 1、8、20 和第 28 个季度为代表，可表示为 $\tau_1 = 1, \tau_2 = 8, \tau_3 = 20, \tau_4 = 28$，产品 X 的表示方法如式(11.10)所示。

$$\text{Min} \sum_{i=1}^{4} \sum_{t=1}^{36} [s_i(t) - \hat{s}_i(t) \times \alpha_t \times \exp(g_t)]^2$$

限制于

$$\left.\begin{aligned}
&s_1(t) = M_1 f_1(t)[1 - f_2(t)] \\
&s_2(t) = f_2(t)[M_2 + M_1 f_1(t)][1 - f_3(t)] \\
&s_3(t) = f_3(t)[M_3 + f_2(t)[M_2 + M_1 f_1(t)]][1 - f_4(t)] \\
&s_4(t) = f_4(t)[M_4 + f_3(t)[M_3 + f_2(t)[M_2 + f_1(t)M_1]]] \\
&f_i(t) = F_i(t) - F_i(t-1) \\
&F_1(t) = \begin{cases} (1 - e^{-(X_1(t)-X_1(0))(p_1+q_1)})/[(q_1/p_1)e^{-(X_1(t)-X_1(0))(p_1+q_1)} + 1], & t \geqslant 1 \\ 0, & t < 1 \end{cases} \\
&F_2(t) = \begin{cases} (1 - e^{-(X_2(t)-X_2(0))(p_2+q_2)})/[(q_2/p_2)e^{-(X_2(t)-X_2(0))(p_2+q_2)} + 1], & t \geqslant 8 \\ 0, & t < 8 \end{cases} \\
&F_3(t) = \begin{cases} (1 - e^{-(X_3(t)-X_3(0))(p_3+q_3)})/[(q_3/p_3)e^{-(X_3(t)-X_3(0))(p_3+q_3)} + 1], & t \geqslant 20 \\ 0, & t < 20 \end{cases} \\
&F_4(t) = \begin{cases} (1 - e^{-(X_4(t)-X_4(0))(p_4+q_4)})/[(q_4/p_4)e^{-(X_4(t)-X_4(0))(p_4+q_4)} + 1], & t \geqslant 28 \\ 0, & t < 28 \end{cases} \\
&X_1(t) = (t - 1 + 1) + \ln(pr_1(t)/pr_1(0))\beta \\
&X_2(t) = (t - 8 + 1) + \ln(pr_2(t)/pr_2(0))\beta \\
&X_3(t) = (t - 20 + 1) + \ln(pr_3(t)/pr_3(0))\beta \\
&X_4(t) = (t - 28 + 1) + \ln(pr_4(t)/pr_4(0))\beta \\
&\alpha_1 = 1, \quad \alpha_t = \alpha_{t-4}, \quad \alpha_t \geqslant 0
\end{aligned}\right\} \quad (11.10)$$

其中，g_t：常数，$0<p_i<1$

$$M_i > 0$$
$$0 < q_i < 1$$
$$\forall i = 1,2,3,4$$

2. 参数估计

NLS参数估计分析结果发现产品X的创新系数与模仿系数在四个世代之间会有显著变动，表11.14列出Norton & Bass模型及Islam & Meade模型的估计参数值，并显示估计参数的意义及在不同世代下的判定系数(R^2)。

表11.14 Norton & Bass与Islam & Meade模型的估计结果

参 数	Norton & Bass	p-value	Islam & Meade	p-value
p_1	0.008	<0.001	0.009	0.006
p_2			0.007	0.000
p_3			0.002	0.005
p_4			0.001	0.007

续表

参　数	Norton & Bass	p-value	Islam & Meade	p-value
q_1	0.324	<0.001	0.415	<0.001
q_2			0.256	<0.001
q_3			0.559	<0.001
q_4			0.813	<0.001
m_1	2128.904	<0.001	1953.800	<0.001
m_2	3990.213	<0.001	4764.968	<0.001
m_3	3500.390	<0.001	2788.022	<0.001
m_4	4331.775	<0.001	1644.746	<0.001
R^2 (1st generation)	0.681		0.816	
R^2 (2nd generation)	0.666		0.815	
R^2 (3rd generation)	0.759		0.917	
R^2 (4th generation)	0.868		0.956	

SMPRT 模型必须估计 16 个参数，包括创新系数(p_1，p_2，p_3，p_4)、模仿系数(q_1，q_2，q_3，q_4)、市场潜力(M_1，M_2，M_3，M_4)、季节性因素(α_2，α_3，α_4)和价格有效性(β)，如表 11.15 所示。

在每种技术下，创新系数均小于模仿系数，在新技术下，创新系数越来越小而模仿系数会越来越大，但预期的总体市场潜力参数估计量为正值，价格有效性为负值，季节性因素表示出今年第一季的影响大于其他季。

表 11.15　SMPRT 模型的参数估计

模型 / 参数	SMPRT 模型	p-value	模型 / 参数	SMPRT 模型	p-value
p_1	0.005	0.001	M_1	2190.823	<0.001
p_2	0.003	<0.001	M_2	5006.861	<0.001
p_3	0.001	<0.001	M_3	3201.802	<0.001
p_4	0.002	<0.001	M_4	1506.606	<0.001
q_1	0.257	<0.001	α_2	0.845	<0.001
q_2	0.196	<0.001	α_3	0.800	<0.001
q_3	0.385	<0.001	α_4	0.830	<0.001
q_4	0.585	<0.001	β	−12.982	<0.001

3. 产品需求预测

个案中共有 36 季的历史数据，以前 34 个季度为训练组来估计 SMPRT 模型的参数，并

利用 SMPRT 预测后两季的产品需求量，然后比较预测结果和实际结果有何不同，预测结果如表 11.16 所示。

表 11.16 一期预测与两期预测结果比较

	一期预测 APE	一期预测差异	两期预测 APE	两期预测差异	MAPE
技术 A	97.1%	8.380	98%	9.121	97.55%
技术 B	4.6%	1.560	8.1%	−1.542	6.35%
技术 C	4.6%	−1.312	26.6%	−7.127	15.6%
技术 D	0.1%	−0.457	21.7%	61.953	10.9%
总计	2.0%	8.172	18%	62.405	10%

在扩散模型中，创新系数和模仿系数会影响产品生命周期的形状，在缺乏历史数据下，以同类商品和决策者的判断推定参数，然而，SMPRT 模型比 Norton & Bass 模型更加复杂，如果要计算高峰时间(T^*)，必须知道 p、q 的值和产品价格。为了降低复杂性，以 Islam & Meade 模型跨世代不同的 p_i 与 q_i 来估计参数。定义 K 在高峰期之前的销售总额(M)，$S(T^*)$则代表高峰期的销售，如式(11.11)所示。

$$\left.\begin{aligned} T^* &= \frac{1}{p+q}\ln\left(\frac{q}{p}\right) \\ K &= \frac{1}{2}-\frac{p}{2q} \\ S(T^*) &= \frac{M(p+q)^2}{4q} \end{aligned}\right\} \tag{11.11}$$

结果如表 11.17、表 11.18 和图 11.6 所示。在表 11.17 中，可以发现估计值和实际值结果相近，显示技术扩散模型可用以预测高峰时间。

表 11.17 高峰期间的估计销售和实际销售的比较

	估计值 T^*	实际值 T^*	K
技术 A	8.997	9	0.489
技术 B	13.569	11	0.486
技术 C	10.185	10	0.498
技术 D	7.747	8	0.499

表 11.18 估计值 $S(T^*)$和实际值 $S(T^*)$的比较

	估计值 $S(T^*)$	实际值 $S(T^*)$
技术 A	211.840	226.274
技术 B	322.307	414.778
技术 C	391.947	422.616
技术 D	335.463	331.389

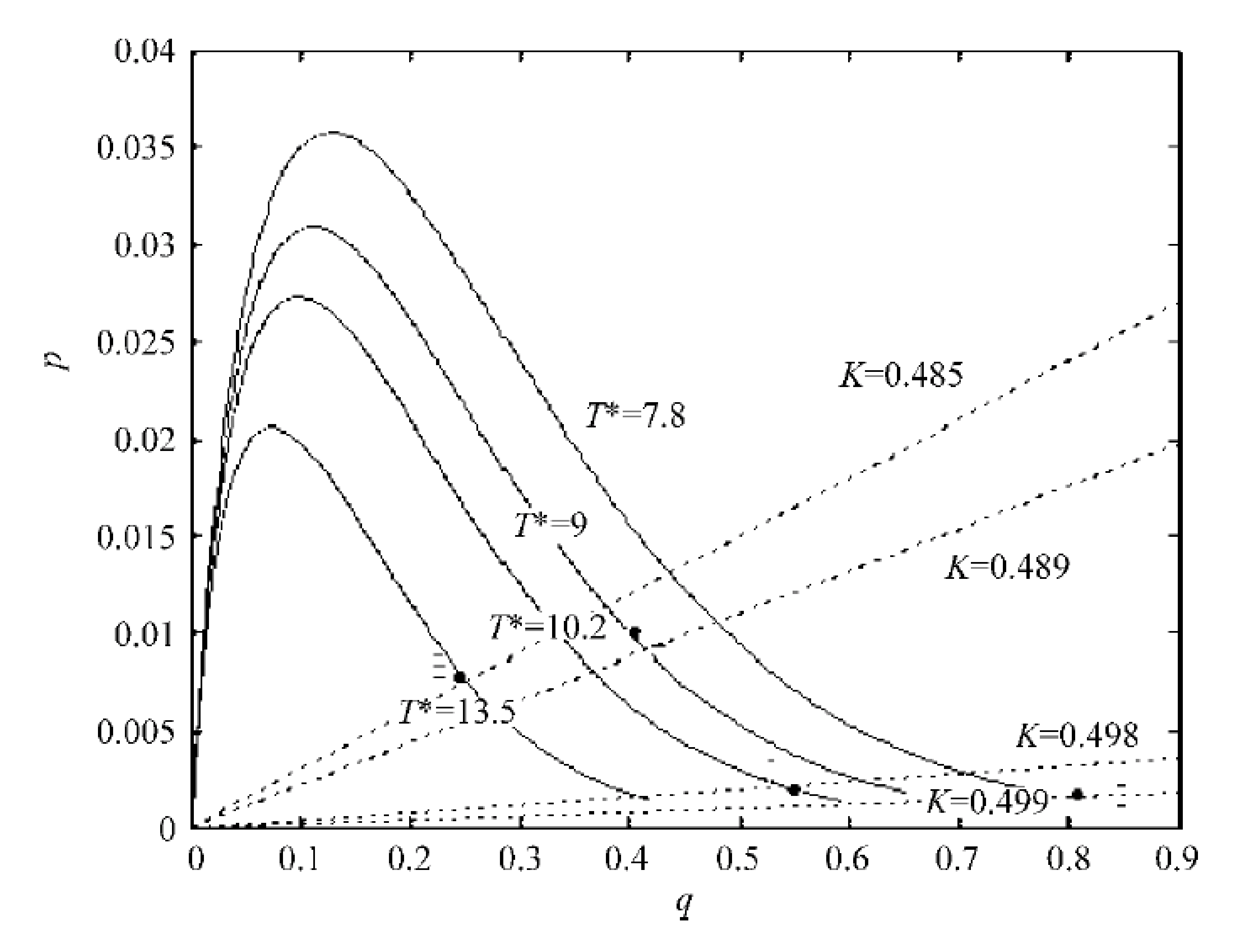


图 11.6　创新系数和模仿系数间的关系

11.5.3　案例小结

本案例提出一个产品需求预测模型(SMPRT 模型)将替代技术、重复购买、价格、市场成长率和季节性因素纳入模型中,并使用非线性最小二乘法估计参数,以实际数据验证此模型,提供需求估计的信息以协助生产决策和规划。

产品的生命周期和需求预测有助于公司在不确定的风险下计划策略。需求预测的结果可以解释和分析领域专家的讨论,以找到解决问题的最佳模式。得到参数的估计值后,即可得出完整的扩散模型,据以描述产品的生命周期和产品需求预测的模式。根据统计分析的结果,判定系数(R^2)代表模型的解释能力,越高代表扩散模型对产品需求的解释能力越高。

11.6　结论

大多数的企业都拥有自己的信息系统,企业会借由各式各样的信息系统累积丰富的企业内部专属数据。若是经理人或决策者无法有效地分析和运用这些数据,而只是一味地储存数据,则此大量数据就无法成为企业的资产,反而成为一种负担。商业智能强调除了利用多种信息技术进行数据的搜集与储存外,更要进一步提供信息分析与报表,并且以便捷的信息存取方式,将有组织以及有价值的企业专属信息转成为有效的决策参考信息,让企业决策者可以增进决策效率以及改善决策的质量。

数据挖掘发掘到的规则可以纳入系统的知识库中不断累积新的知识,结合公司内部所拥有的领域与专业知识以建立起知识体系,并纳入企业的知识管理系统,最后再结合企业决策者自身所拥有的经验以及能力,将知识灵活应用,进而成为企业专属的智能。

问题与讨论

1. 请比较商业智能的不同定义，并讨论各个顾问公司如 IBM 和 Gartner 对“商业智能”的观点的异同。

2. 请搜集商业智能的方法和系统的应用案例，并加以讨论。

3. 推荐系统(recommender system)的应用近年来越来越广泛，主要根据搜集的信息，预测顾客的喜好进而提供顾客需要的商品，请举一个应用案例，并试着说明如何建议其分类或预测模型。

4. 承上题，协同过滤(collaborative filtering)是推荐系统上重要的功能，请说明在电子商务或网络购物中，如何根据在线事务数据提供顾客推荐商品。

5. 若你是一位信用卡发行公司的主管，手中握有庞大的顾客事务数据，请思考你将如何应用数据挖掘与大数据分析的技术，并说明可能的应用方向。

6. 附件数据 Process. csv(请于本页二维码中下载)为某加工制程所搜集到的三项反应值与加工时间，即 x1、x2、x3 与 date，并且该制程于时间 2011-04-25 18:00 与 2011-04-25 23:00 时有机械故障问题发生。根据以上情况，请问工程师该如何由数据中预测未来同类型故障的发生?

7. 附件数据 Osteoporosis. csv(请于本页二维码中下载)为针对骨质疏松症研究所搜集的 1000 笔数据，其中包含“年龄”、“性别”、“血型”、“家族遗传”、“骨质疏松”等特征数据。假设欲通过 CART 了解“年龄高于或等于 50”、“性别”、“血型为 O 型”与“有无家族遗传”对骨质疏松的影响。请回答下列问题：

(1) 请问应如何对此数据进行数据预处理?

(2) 请进行 CART 的构建，并归纳对骨质疏松可能的影响因子。

(3) 请由分析结果层别骨质疏松发生的高风险族群，并对骨质疏松的防范进行建议。

8. MAE、MSE 与 MAPE 为三种不同的误差评估方式。请分别回答下列问题：

(1) 请问 MSE 与 MAE 何者对离群值较为敏感?

(2) 假设模式 A 为某灯泡故障时间的预测模式，模式 B 为某日光灯故障时间的预测模式。请问若要比较模式 A 与模式 B 的预测误差时，MSE、MAE、MAPE 何者相对较为恰当?请说明原因。

(3) 假设要以一度量来呈现手表使用一年后时间的误差时，MAE 与 MAPE 何者相对较为恰当? 请说明原因。

(4) 假设以 $y=5+3x_1-2x_2-x_3$ 为附件数据 Reg. csv(请于本页二维码中下载)的预测模式，请分别计算此模式对于数据的 MSE、MAE 与 MAPE。



第 12 章

制造智能

12.1 序言

在智能化与自动化的制造环境下，巨量数据在生产过程中被自动或半自动地记录和储存在工程数据库等相关数据库中，这些巨量数据究竟是资产还是负债，取决于其数据价值发挥与否。**制造智能(manufacturing intelligence，MI)**是整合数据挖掘工具、大数据分析方法及自动化系统以建立智能化制造系统，以探索和分析大量制造数据，发掘潜在有用的信息、有意义的样型或规则等，作为提升产品良率、增加生产力、动态规划产能、优化制造资源分配以及降低生产周期时间等制造决策的依据。

在半导体等高科技产业的制造过程中，产品本身的数据、所用的制程技术、配方及经过的加工机台，大量的制程数据和产品在生产过程中经过机台加工产生的工程数据，或是为了监控产品质量与制程的稳定性、故障分析，而以人工输入方式记录的数据来进行制程监控的数据，以及制造管理的信息，都会被自动搜集记录在各种数据库中。

半导体纳米制程的技术难度和变异有增无减，完全自动化的 12 英寸晶圆厂月产能超过十万片，在线同时用十几种制程配方参数(recipe)生产各种产品，每片晶圆要经过数百道至上千道反复循环的制造程序，每个工作站有几个到几十个精密的反应室(chamber)可以选择，生产过程中可以随着时间读取几万种实时监控数据、近万个在线抽样检测的量测值(metrology)和几百种在一片晶圆上不同位置测量的电性测试参数，平均每片晶圆上可以读到的相关数据就超过百万笔以上，再加上集成电路复杂的生产模式，使得数据除了具有大数据常见的 4V(volume，variety，velocity，veracity)特性之外，还有数据主效应不明显、数据分布不均衡、前后制程的交互作用复杂等挑战(简祯富，2014a)。

以半导体制造为例，主要包括以下形态的数据：

(1) **生产(production)数据**：每片晶圆在制造过程中的描述数据。例如，货批编号、产品名称、产品通过站别、产品通过站别的日期与时间、产品通过站别所使用的机台名称等。

(2) **测量(metrology)数据**：针对某一批货所搜集的数据。例如，产品测量参数名称、测量的时间、测量机台、产品测量参数值、测量参数规格上限与下限等。

(3) **设备(equipment)数据**：针对某一机台状况所搜集的数据，通常会跟随着生产过程和预防保养来进行搜集。例如，机台监控参数名称、机台参数值、机台参数规格上限与下限、预防保养和维修记录等。

(4) **缺陷(defect)数据**：描述产品缺陷状况的数据，通常来自于监控设备(inspection equipment)的记录、故障分析(failure analysis)、特征分析(signature analysis)等。例如，缺陷的层别名称或编码、每层缺陷个数、缺陷密度、每芯片缺陷个数等。

(5) **晶圆允收测试(wafer acceptance test，WAT)数据**：晶圆通过电子特性测试(E-test)的结果。例如，每批货的芯片数、测试电性参数名称、测试电性参数值、电性测试的时间、所用的测试设备、测试电性参数的规格上限与下限等。

(6) **电性功能针测(circuit probe test，CP)数据**：每颗晶粒(die)探针测试后的结果。例如，晶粒位置、针测结果和 Bin 值等。晶圆图(wafer bin map，WBM)上故障晶粒的分布可分成三种：

① **随机性故障(random defect)**：指故障晶粒没有一定的样型或群聚，而是随机分布在晶圆上。随机性故障的产生很难完全消除，例如制造过程中的微尘(particle)所造成的故障。

② **系统性故障(systematic defect)**：指同一批晶圆中，故障晶粒因为特殊原因导致特殊的晶圆图形，例如环状、边缘不良、棋盘状等，如图 12.1 所示。因此，可借由分析故障晶粒(fail die)所呈现的空间分布追查可能发生问题的制程或是机台，如显影时光罩对不准(photo-mask misalignment)、过度蚀刻(over etching)等。系统性故障产生的原因通常有迹可寻，因此只要找出造成系统性故障的样型即可推测出异常的原因，进而消除这些系统性故障。

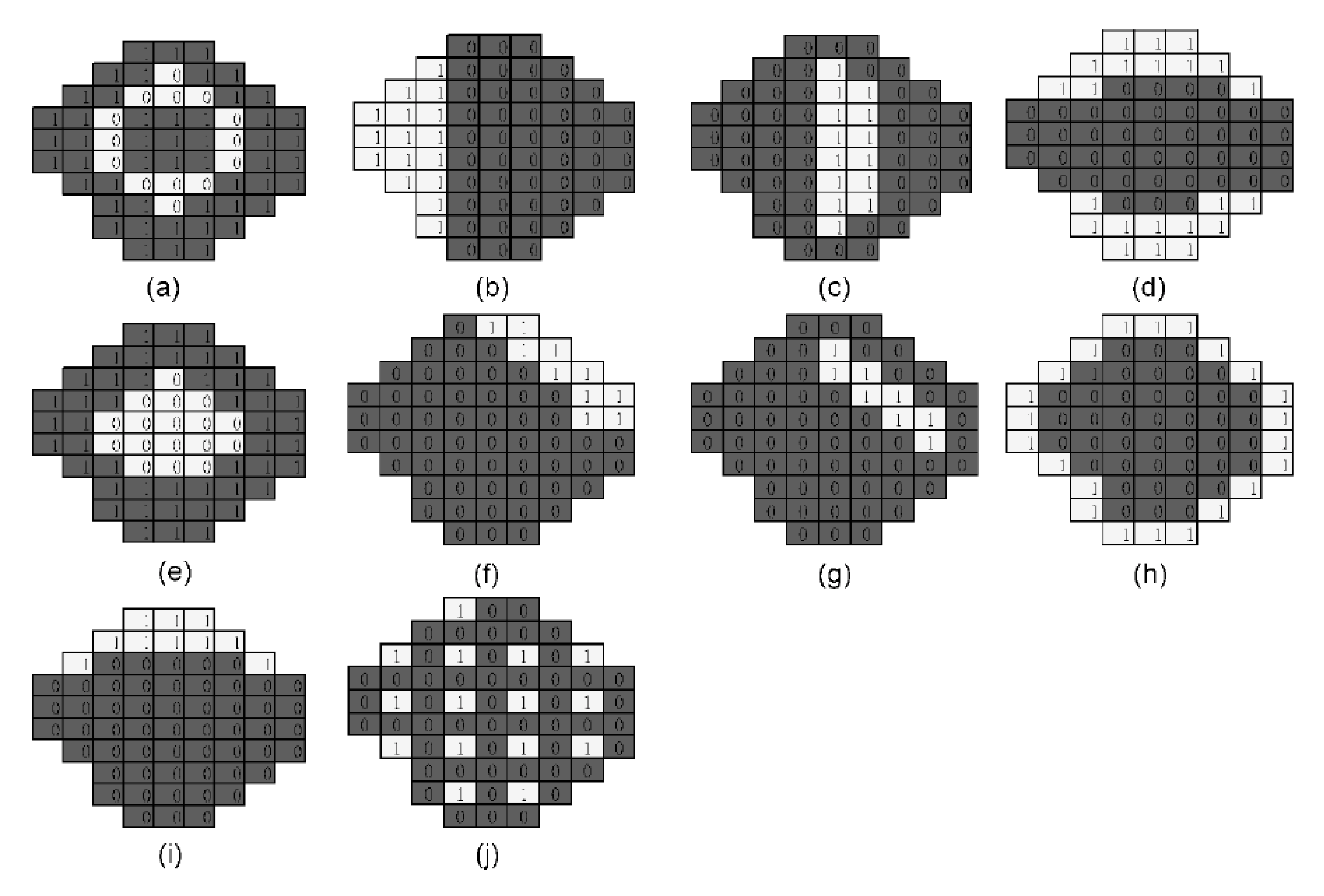


图 12.1 系统性故障

③ **混合型故障(mixed defect)**：指同时有随机性故障与系统性故障而产生的晶圆图，一般工厂常见的晶圆图多半属于此类型，如图 12.2 所示，因此必须从随机性故障所造成的噪声中，提取其中较易移除的系统性故障样型与原因。

随着半导体制程持续微缩挑战物理极限，允差也不断紧缩，使得即使是资深工程师也很难单凭专业知识和经验法则(rule of thumb)或传统的统计分析方法，从巨量数据中迅速找出制程异常的原因，以减少产品报废损失。例如，同时经过沉积制程的某机台与蚀刻的某机台后，特别容易造成芯片良率过低，或是哪些制程测量参数值倾向于一起变动，或是某制程

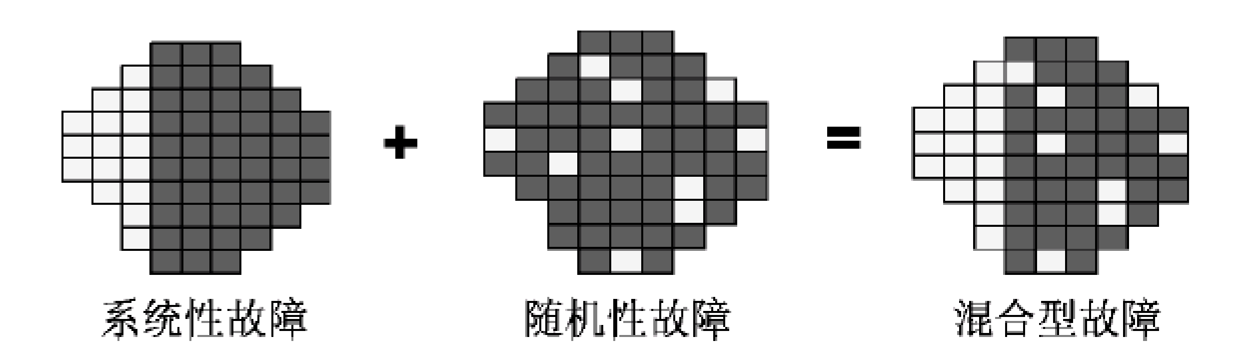


图 12.2　混合型故障

的主作用项影响不显著,但制程间的交互作用项却造成严重影响,或是某机台在某段时间的表现较差等。尽管商用的统计软件逐渐可以支持大数据分析,但是由于缺乏针对半导体产业需求和特性的应用模块,影响了一般工程师的使用意愿。

根据本书的数据挖掘架构,利用半导体制造数据和各种实际案例,说明实际应用大数据分析和数据挖掘方法,以提取制造智能,有效协助工程师在短时间内缩小范围,找出造成事故问题的真正原因,作为工程师及领域专家解决问题的参考依据。

12.2　WAT 参数特征提取与关联分析

1221　案例说明

半导体内存组件可分为挥发性与非挥发性两种。挥发性是维持、保有内存内的数据须依赖持续的电源供应,如动态随机存取内存(DRAM)与静态随机存取内存(SRAM);反之,非挥发性内存即使遇到电源中断,其内部存储器的数据仍得以保存,如 EPROM、罩幕式只读存储器(mask ROM)及闪存(flash)。

在完成所有晶圆加工步骤后,都会在制程结束前进行晶圆允收测试(WAT),或称电子特性测试,以测试半导体组件上的电子特性,而每一个参数都是用来监控组件的某个特性,因此往往会与特定一层或多层的制程特性有关。例如,某起始电压(voltage,V)过高,多半是因为在制造此组件时离子植入掺杂值偏高,所以借由电性测试结果即可诊断晶圆发生异常的原因。电性测试的参数往往超过上百项,针对不同需要,测量不同的电子特性,如电阻、电压、电流、电感等。目前半导体厂的作法是抽测每一批(lot)生产的晶圆,一批晶圆抽 5 片,每片测 5 点。此外,由于每个电性参数都有既定的规格,所以测量的数据需与规格作比对,以监控产品质量。

以往半导体的事故诊断主要依靠工程师的领域知识,或层别制程站别或机台的差异来找出可能发生变异来源的机台(简祯富等,2001),对于事故诊断大多仅借助统计制程品管或采用无母数检定比较其参数或机台表现差异。另一方面,由于数据维度与数据数量越来越大,变量之间复杂的交互作用,加上不同数据搜集来源混杂的噪声,传统统计分析方法有其限制,因此必须借助数据挖掘和大数据分析技术。本案例通过多变量的群聚分析技术,根据多维度属性予以划分为不同群聚,并选择合适的规则以归纳描述对应的特征,提供工程师作为事故诊断的参考。

本案例(简祯富等,2003)是针对半导体制程事故诊断的数据进行特征提取与描述,通过人工神经网络的自组织映射图算法先将半导体晶圆允收测试数据分群,以发现隐藏于数据中的样型与良率间的关联性,了解参数表现的概况(profiles)与良率间的关系,再用决策树

将良率异常类别的特征以树状结构呈现，并转换为分类规则，提供工程师作为监控制程变化与事故诊断的决策依据。

1222 分析过程

1. 数据准备

本案例以某半导体厂的实际数据为实证。此公司是集成电路研发、制造、测试及销售专业厂商，专注于非挥发性内存(non-volatile memory)及系统整合芯片 IC 产品，为全球非挥发性内存主要供货商。

针对搜集的多维度数据，首要工作是数据准备与探索。由于要挖掘的是 WAT 数据与良率间具特殊分布的样型。数据字段包含每批晶圆的批号、测试时间及各测试参数的测量记录。由于测量变量众多且皆为连续型变量，在与工程师讨论之后选择此项产品的 41 个主要参数作分析，共取得 264 笔数据。在运算前，对每个变量进行归一化转换的前置处理。

本案例先应用 SOM 神经网络进行聚类分析，以发现特殊数据样型与良率间的关联性，发现低良率的特征。先建立两层前向连接的神经网络，将高维度的图样特征，映射至二维的输出神经元数组。通过对特殊样型的观察，定义欲区别的“群别”作为决策树分类的目标。通过 SOM 神经网络可以同时考虑多变量的因子，甚至察觉出先前未知的信息，而不需事先局限住可能变因的范围。SOM 拓扑图除了可以展现数据之间的群聚关系外，亦可通过检查个别变量的分群状况以颜色区分各变量对于特定聚类的贡献程度。检查拓扑图的聚类分布后，引入与良率关联的相关变量，以作为探讨其各聚类对于相关变量的分布后，将要划分的聚类新增一类别字段，以作为后段决策树分类的目标(target)变量。

了解数据特性与分布后，设定 SOM 输出节点数为 1000，并将数据输入 SOM 网络，通过向量量化与向量投影，其群聚结果如图 12.3 所示，在了解其数据点在拓扑图上的分布后以颜色来区分群聚，共可分为四群。根据上述分群方式针对这些数据点的良率值进一步分析(如图 12.4 所示)，其中良率值的相对表现越高，颜色越接近红色(如群 3、群 4)，反之则接近蓝色(如群 1、群 2)。由良率分布可以发现右下角的群聚其良率相较于另外三群较低，大多在 0.75 以下。针对这样特殊的样型，在与领域工程师讨论并整理过往的诊断记录后，引入拓扑图中。

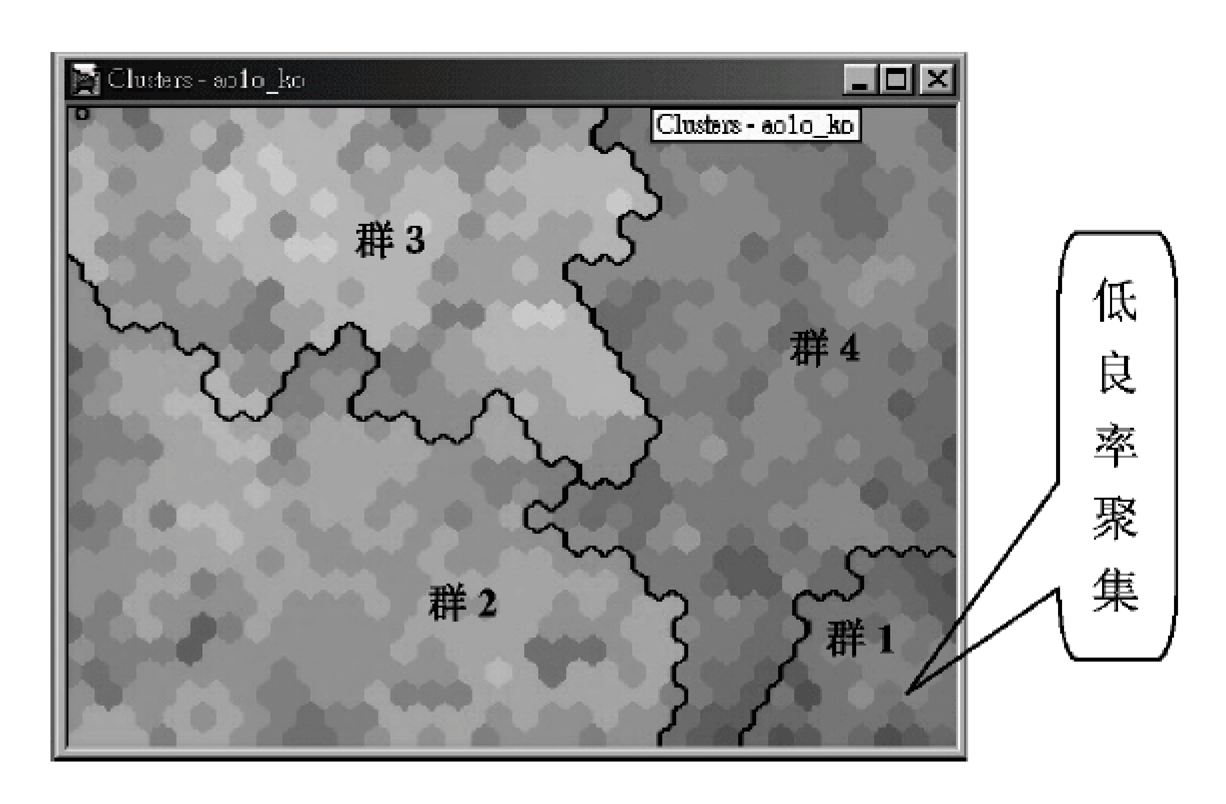


图 12.3　WAT 数据群聚现象拓扑图

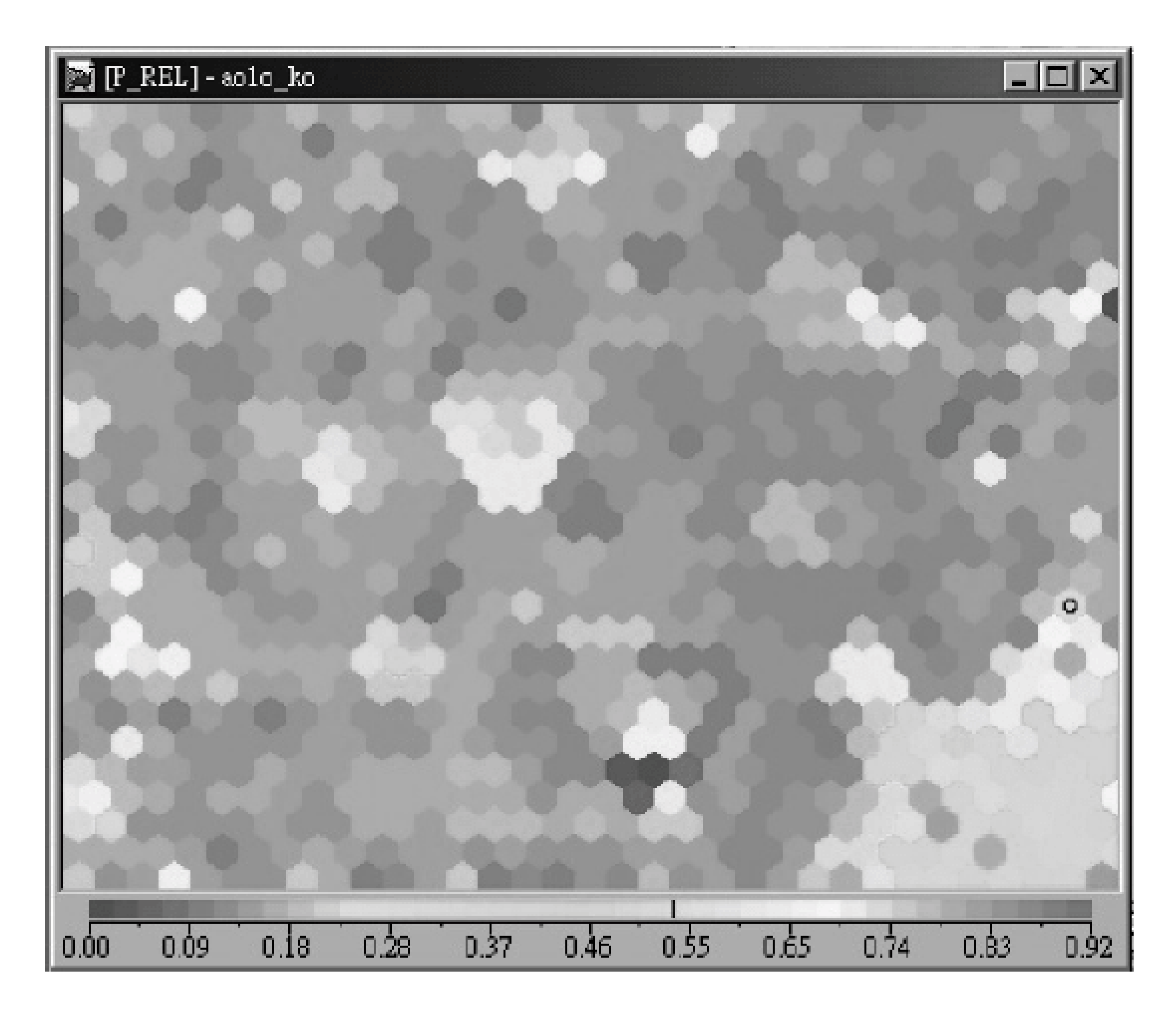


图 12.4 良率分布拓扑图

发现群聚位在拓扑图右下角的数据点皆被工程师下过相同的诊断记录(代号皆为“#”)且对应图 12.5 皆影响到良率的表现。由于群聚样型是根据 WAT 参数的表现而分群,因此群 1 的特征与低良率现象可能存有某种关联。通过检查各变量对群聚现象的贡献程度,拓扑图亦可找出哪些参数对于群聚及良率有较大的贡献。但由于 SOM 群聚分析着重以可视化方式表现群聚,因此接着以决策树进行特征提取与分类规则的描述。

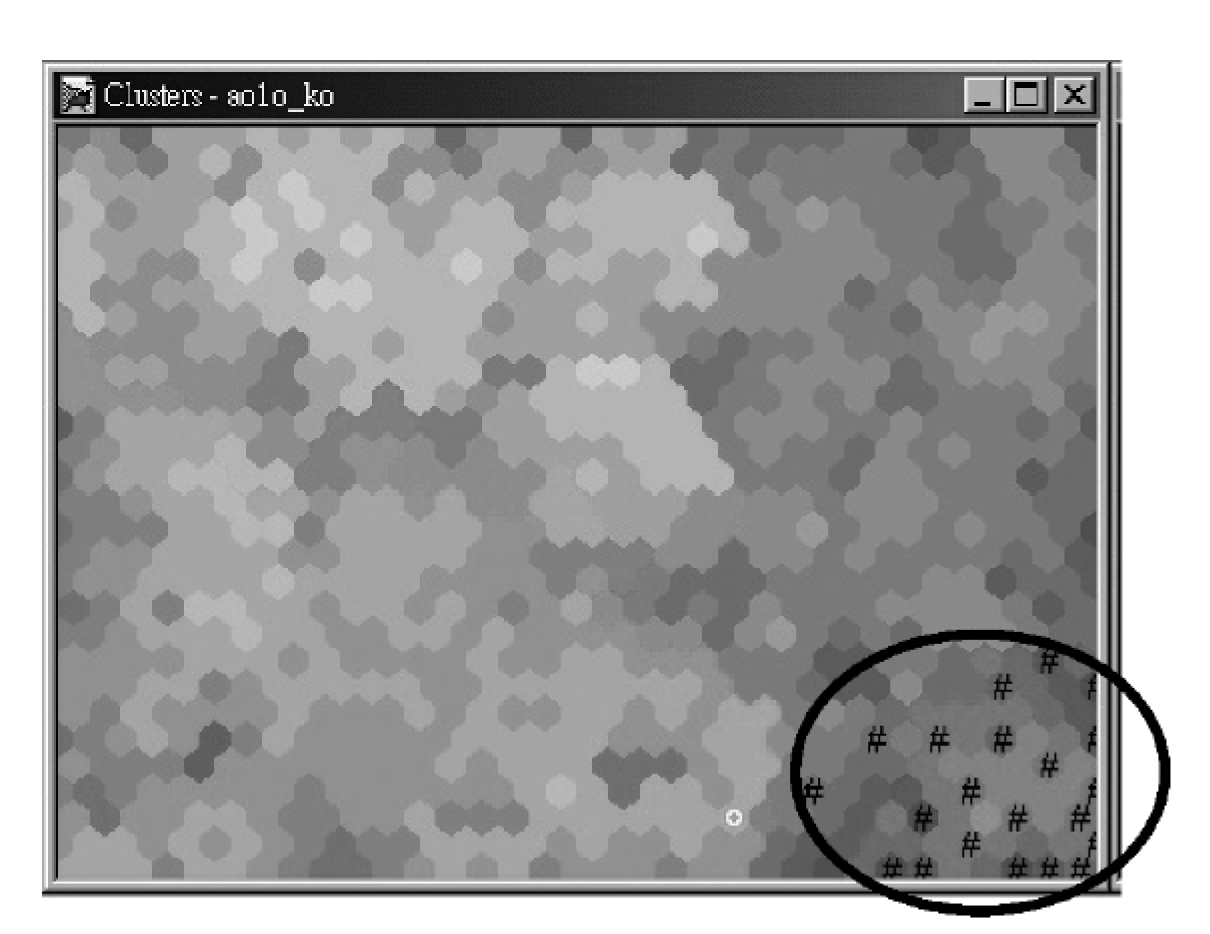


图 12.5 右下群聚的诊断记录

2. 建立数据挖掘模式

本案例以决策树模式提取各分群的特征,通过参数表现特征提供给工程师监控制程变化的决策依据。决策树以树枝状架构呈现其分类结果,其中指向同一分群的规则可视为其样型特征。SOM 分群中发现位于拓扑图右下方的群 1 有相同的事故记录,因此以群 1 为目标将其类别定为“Bad”(共 21 笔),其余三群的标签则定为“Other”(共 243 笔)。利用决策树进行分类,而 WAT 参数表现的差异将群 1(Bad)与其他(Other)划分。

经由决策树叶节点至根节点可产生不同的分类规则,以 Gini 系数衡量分类节点的纯

度，Gini 系数越小代表该节点的类别纯度越高，而 Bad 类别的正确率代表节点中原有所有 Bad 类别个数中经由该规则能被正确划分的比例。换言之，期望找到 Gini 系数与正确率高的规则来代表聚类特征，决策树的分类结果如图 12.6 所示，发现借由变量 *V*9 与变量 *V*18 即可将 Bad 类别与 Other 划分且正确率可达 90.4%(19/21)。若以分类规则的纯度衡量，当 *V*9<8.59 且 *V*18≥−7.8 时，其规则的 Gini 系数为 0.095。当然规则的纯度越高越好，但由于本案例关心的是群 1 所显现出的特征差异，因此其他规则虽然纯度很高，所能代表群 1 特征的数据点数却很稀少。其中，“*V*9<8.59&*V*18≥−7.8”代表的是群 1 与其他群聚间差异的特征，由此推断低良率与特定的量测参数是高度相关。另一方面，通过特定表征与事故诊断连接的规则库，可以提供工程师于制程的监控、分析与良率预测的参考。

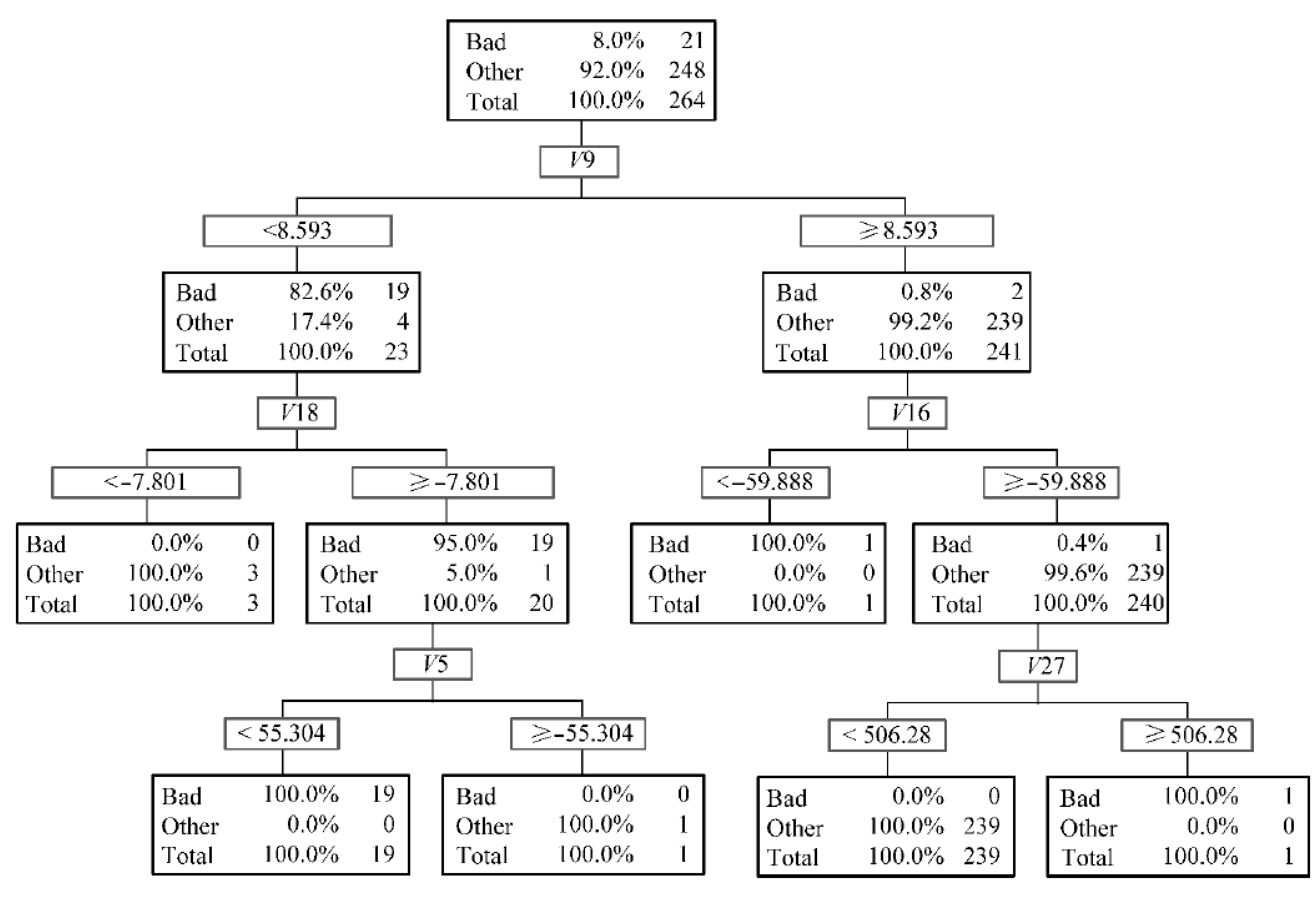


图 12.6 决策树分类结果(简祯富等，2003)

3. 结果诠释与评估

使用决策树分类时最常碰到数据变量间具共线性或高度相关的情况。本案例在分支节点检查以卡方检定分析各变量在此分支点对于分类目标变量的贡献程度，为了显示上将检定结果的 *p*-value 取对数函数转换为 −log(*p*-value)，−log(*p*-value)值越大表示其贡献程度越高。在第一层决策树分割的阶段其各变量的−log(*p*-value)如表 12.1 所示。

表 12.1 各变量贡献程度(因篇幅限制只列出前几项)

贡献排序	变量	−log(*p*-value)	贡献排序	变量	−log(*p*-value)
1	*V*9	41.74	3	*V*11	38.14
2	*V*22	40.83	4	*V*19	37.60

续表

贡献排序	变量	$-\log(p\text{-value})$	贡献排序	变量	$-\log(p\text{-value})$
5	*V*18	37.57	13	*V*16	13.24
6	*V*14	34.69	14	*V*36	11.14
7	*V*4	34.66	15	*V*26	10.51
8	*V*10	30.96	16	*V*34	10.27
9	*V*20	25.12	17	*V*12	9.44
10	*V*35	23.21	18	*V*32	7.48
11	*V*15	21.03	19	*V*33	7.33
12	*V*23	20.93	⋮	⋮	⋮

检查各变量的贡献程度可以发现，在第一次分割时虽然变量 *V*9 仍较变量 *V*22 贡献大，但两者贡献相当接近。有鉴于此，除了第一次以变量 *V*9 作切割提取特征规则以代表群 1 样型表征外，另外选取其他贡献程度前九名的变量（*V*22～*V*35），各自进行分割以讨论其分类结果，各项选取分类的结果规则如表 12.2 所示。

表 12.2 贡献程度前九名分类结果

Rule	Description	Bad(21)	Other(243)	Bad 正确率	Gini 系数
1	$V22\geqslant -7.99$	18	3	0.857	0.245
2	$V11<-8.27$	17	3	0.810	0.255
3	$V19<-1.83$ & $V35\geqslant 7.87$	20	1	0.952	0.091
4	$V18\geqslant -7.764$	16	2	0.762	0.198
5	$V14<8.69$ & $V10\geqslant 3.46$	18	0	0.857	0.000
6	$V4<8.47$ & $V10\geqslant 3.46$	18	0	0.857	0.000
7	$V10\geqslant 3.58$ & $V4<8.48$	17	0	0.810	0.000
8	$V20\geqslant 58.72$ & $V4<8.48$	18	2	0.857	0.180
9	$V35\geqslant 8.06$ & $V4<8.48$	19	1	0.905	0.095

由表 12.1 与表 12.2 可以发现，所选出的前几名变量，皆可以将原本的数据做出划分，但是划分的规则正确率及 Gini 系数皆不相同。以规则 1 为例，其分类的正确率可达 85.71%（18/21），Gini 系数也下降至 0.245；以规则 3 来说，第一次以变量 *V*19 作为分支时，其正确率可高达 95.23%（20/21），但规则 Gini 系数只有 0.287，在引入变量 *V*35 继续分支的情况下，才能在正确率不变的情况下将 Gini 系数降低至 0.091。由于在第一个节点进行分支时变量 *V*19 对于目标分类的显著程度不如变量 *V*9，因此在预设以贡献程度高的变量进行分支的条件下，会以变量 *V*9 进行分支提取规则，而得到“当变量 $V9<8.59$ 且变量 $V18>-7.8$ 时，属于某项会造成低良率的事故原因”。

另一方面，与工程师讨论后发现，群 1 这种 WAT profile 的特殊样型是由于蚀刻过程造

成残余物质,影响部分组件的漏电电压而导致良率下降,针对各变量对于群聚现象的贡献,检查各变量的拓扑图也能发现与量测漏电电压有关的参数表现与良率间的相关性(如图 12.7 所示),其中群 1 在这些变量的表现上相对于其他群皆是迥异的。在决策树所提取出的参数表现特征中,亦可发现变量 *V*9 与变量 *V*18 测量的皆是组件的漏电电压。由此可以验证在群聚现象中所发现的样型,借由特征的提取有助于建立 WAT 参数表现与低良率之间的特征规则。

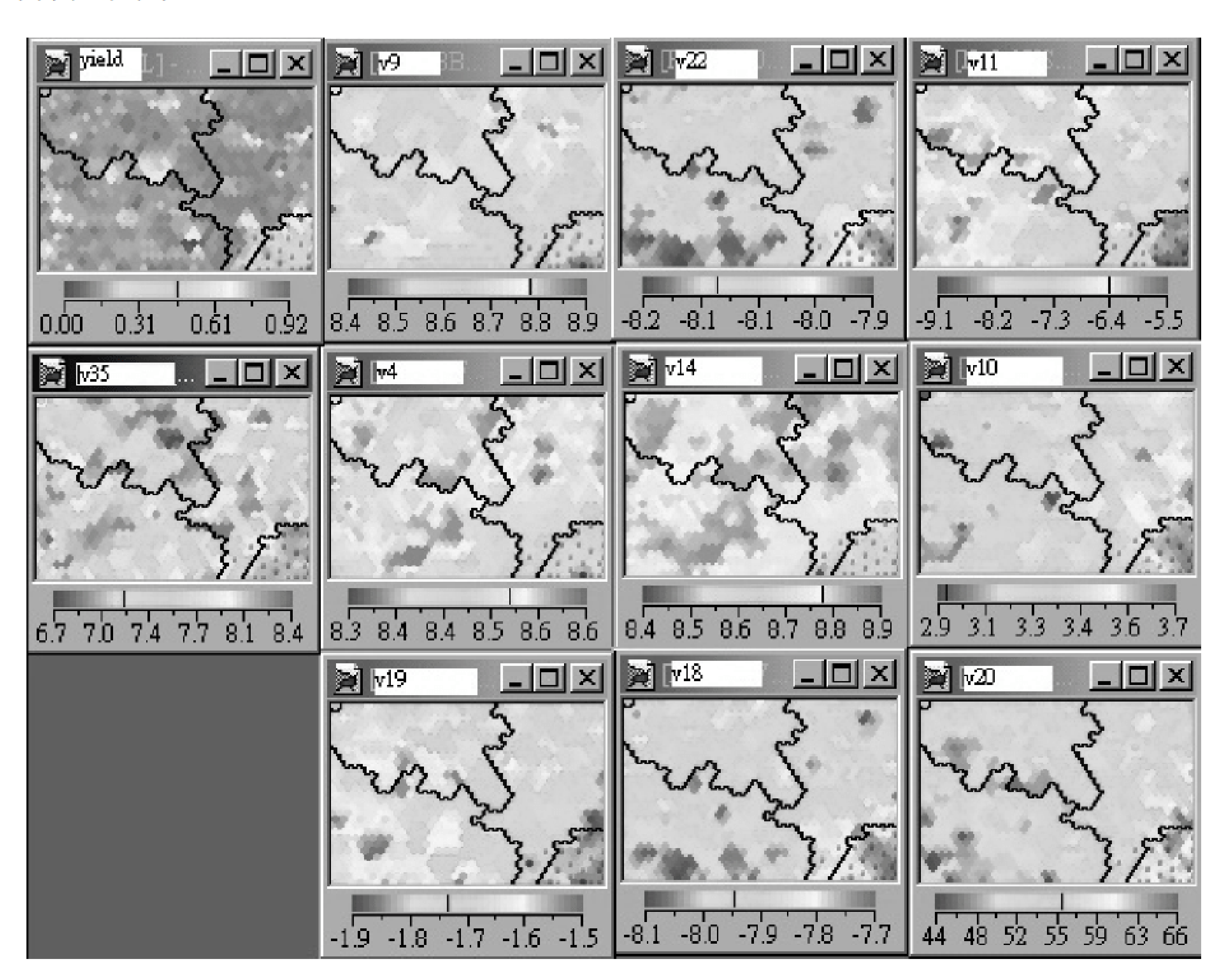


图 12.7　各群聚在测量漏电电压相关变量的相对表现(简祯富等,2003)

12.2.3　案例小结

本个案利用自组织映射图网络算法及决策树分类规则,经过群聚分析找出与良率分布相关的样型,并通过特征提取与描述表达群聚特征的数据挖掘方法,协助工程师进行半导体产品的监控。未来更可加入其他相关的数据例如制程记录数据、测量数据等,进行多变量关联性分析后借由特征差异的分析与比较,协助工程师进行事故诊断与制程优化,加速判别产品的良率水平及故障类别。

12.3　半导体 CP 测试数据挖掘与晶圆图样型分类

12.3.1　案例背景

半导体晶圆在制造过程中,可能受到制程事故异常因素干扰,因而造成晶粒的 CP 良率过低,使晶圆图出现某些特殊的故障样型,需要工程师尽快厘清问题的根本原因或找到解释

会发生此异常的原因，以进行制程改善避免更多的损失。

半导体针测完成后，所累积大量的晶圆图和相关测试数据，即可提供制程工程师追查制程发生异常问题的线索(Chien *et al*., 2013; Liu & Chien, 2013; Hsu & Chien, 2007; 简祯富等，2002)，例如有问题的机台与发生异常的制程等。然而，实务上大部分仍依赖工程师以人工目视判断的方式来分析晶圆图，因此可能由于人为主观因素及对空间图形辨识能力的差距，造成判断结果的不一致与故障原因分类的人为偏差，以致无法快速排除故障减少损失。

晶圆图是一种显示晶圆上各晶粒检测结果的图形化数据，主要包括缺陷图(defect map)与针测图(bin map)两种，晶圆图是追溯产品异常原因的重要线索，借由晶圆图的模型分析得以找出可能造成低良率的原因。在晶圆制造过程中，最后测试阶段会进行不同电性功能的针测(electrical wafer sort)，以确保产品的功能性。WBM 是晶圆制造测试过程因为不同测试结果所产生的图形。图中以芯片为单位，通常以不同颜色或故障代码(bin code)标示于各个芯片位置上代表测试完成的结果。一般由特定针测结果(故障代码)的空间分布(spatial distribution)情况，可推导出造成此结果的制程原因，故工厂都会记录每片晶圆经过测试后产生的晶圆图，如图 12.8 所示，以作为事故诊断之用。

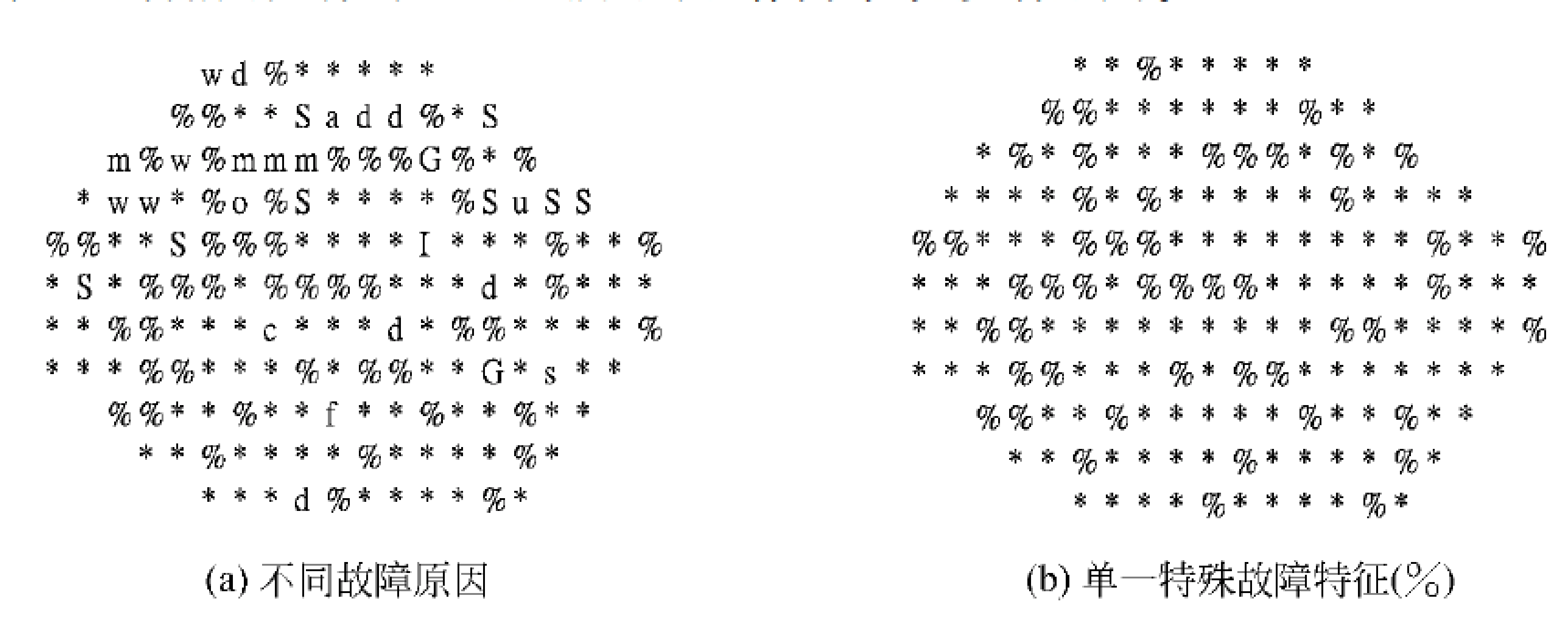


图 12.8　晶圆图示例

本个案结合空间统计检定方法与自适应共振理论(adaptive resonance theory, ART)人工神经网络，发展晶圆图分类的流程和运算法则，将大量且紊乱不一的晶圆图，根据故障晶粒呈现的群聚现象作有效的归群整理，并记录其共同表征(common pattern)，以建立系统化分类的晶圆图库(Hsu & Chien, 2007)。除了晶圆图样型分类外，也提供工程师将发生异常现象的晶圆图与分类过的晶圆图库，进行图形相似度比对。借由寻找与过去亦发生相同群聚现象的图形及其原因，结合领域知识推论其在制程上可能经历的问题再加以验证，缩短工程师故障排除(trouble shooting)的范围与分析所需的时间。

12.3.2　分析过程

1. 数据准备

本研究的晶圆图为某一光罩只读存储器的探针测试数据结果，共 138 批货(lot)，每一片晶圆中，扣除掉特殊位置没有测试外，总共有 268 片芯片有标记最后结果。

工程师可根据不同故障代码，追查导致某种特定样型出现的原因。若是此芯片在前项的电性针测已经发生故障，则这芯片位置就不再做其他针测，而直接标记故障代码，至于通

过所有针测的芯片则会标记为通过，表示此芯片良好。

原始晶圆图为一个二维的文本文件数据，对于视觉判断上需再以图形的方式呈现以利用后续分析与晶圆图浏览比较。在分析前需将原始数据转换为分析所需的数据格式，包括以下两个步骤：

（1）**建立二维图形坐标**：针对故障建立二维图形坐标，若芯片位置为故障晶粒则以 1 表示，非故障晶粒则以 0 表示，如图 12.9 中二元图形所示。

（2）**二维图形坐标转换一维数字向量形态**：由于分析上数据格式的限制，需将原始的二维图形数据，以由左而右由上而下的方式重新编码成一维的二元向量，转换后的一维向量则可直接应用于 ART1 网络计算。

晶圆图数据转换步骤包含两种转化。先依据工程师经验，分析时选择 0.15 为临界值作二分法的判别，若某批货在此位置出现故障比例大于 0.15，则此位置标记为“1”(bad)，反之则标记为“0”(good)。接着，将原始二维坐标，转换成一维的二元向量，其中，芯片位置为良好则以 0 表示，若为故障晶粒则以 1 表示，因此总共产生 138 笔二元值的向量。

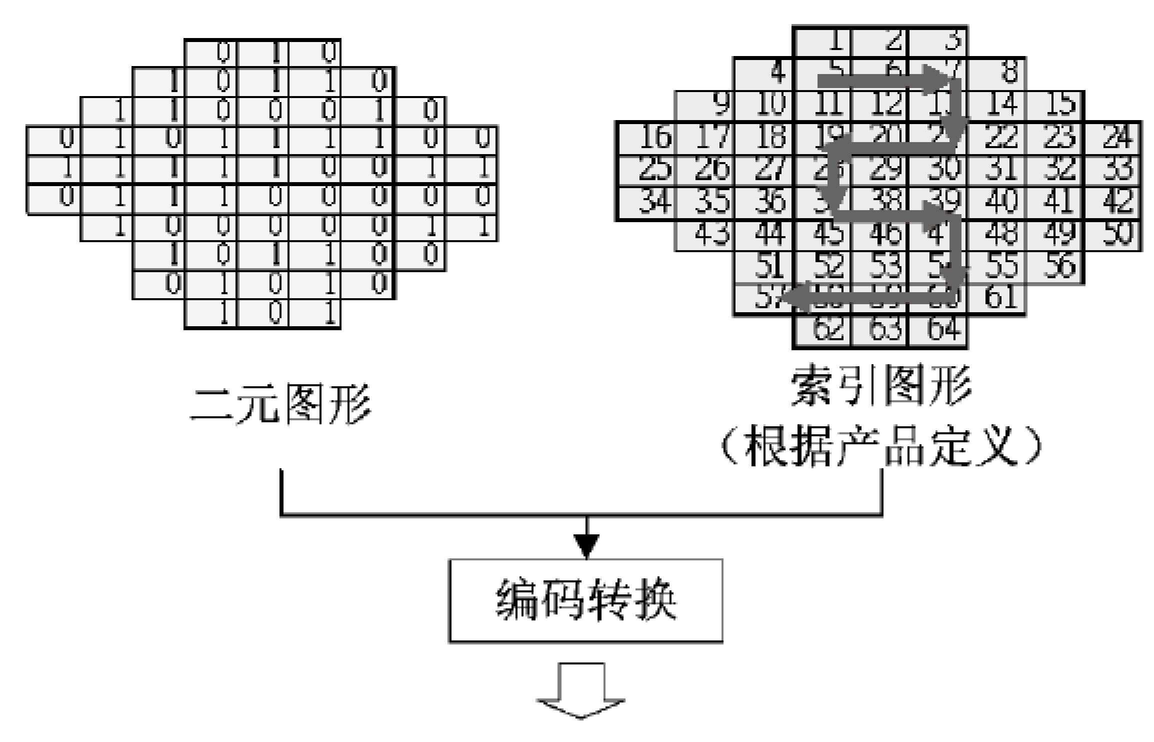


图 12.9 转换二维晶圆图数据成一维向量

2. 空间统计检定

本个案利用空间统计检定方法将晶圆图分成四类：光罩错误的晶圆图、随机性晶圆图、特殊群聚的晶圆图以及其他等四类晶圆图。对于随机性的图形，依其故障严重程度分成两大类，并不再进行图形比对。对于特殊性晶圆，则再利用人工神经网络进行聚类分析并产生一组具有特殊图形的样板，再以此样板与归类于其他性的图形进行 ART1 相似度比对，以便进行分类。

根据加特(Gart)和兹韦福尔(Zweiful)提出的空间统计检定，可检测空间中两类别的数据是否有关联，其修正后的统计量如式(12.1)所示，进一步讨论可参见（Agresti，1990）。

$$\hat{\theta} = \frac{(N_{GG} + 0.5)(N_{BB} + 0.5)}{(N_{BG} + 0.5)(N_{GB} + 0.5)} \tag{12.1}$$

首先定义晶圆图显示状况，在晶粒 i 位置上若出现故障，则标记为 $Y_i = 1$(Bad)，否则视为正常 $Y_i = 0$(Good)。在考虑 King-Move 邻近区域如图 12.10 所示下，可建立如表 12.3 的 2×2 列联表以考虑故障晶粒与正常晶粒在二维空间上的关系，并可计算 N_{GG}，N_{GB}，N_{BG} 及 N_{BB} 等 4 个值。

表 12.3　晶粒相邻位置关系的列联表

位置 j / 位置 i	Good	Bad
Good	N_{GG}	N_{GB}
Bad	N_{BG}	N_{BB}

$$N_{GG} = \sum \sum_{i<j} \delta_{ij}(1-Y_i)(1-Y_j) \tag{12.2}$$

$$N_{GB} = \sum \sum_{i<j} \delta_{ij}(1-Y_i)Y \tag{12.3}$$

$$N_{BG} = \sum \sum_{i<j} \delta_{ij}Y_i(1-Y_j) \tag{12.4}$$

$$N_{BB} = \sum \sum_{i<j} \delta_{ij}Y_iY_j \tag{12.5}$$

其中，$\delta_{ij} = \begin{cases} 1, & Y_i \text{ 和 } Y_j \text{ 在 King-Move 邻近区域(如图 12.10)}, \\ 0, & \text{其他}。 \end{cases}$

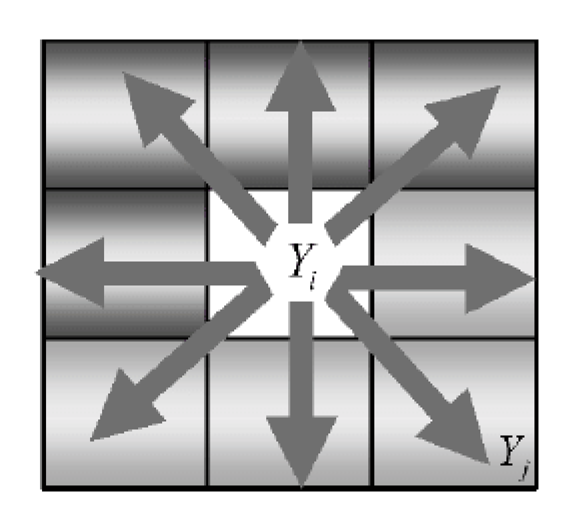


图 12.10　King-Move 邻近区域

本研究所用的空间统计检定步骤如下：

步骤 1：检定假设。

H_0：晶圆图上故障晶粒或正常晶粒呈现随机分布(即无任何特殊群聚或离散的现象)。

H_1：晶圆图上故障晶粒或正常晶粒呈现非随机分布(即发现有特殊群聚或离散现象)。

步骤 2：检定统计量。

在大样本下 $\ln\hat{\theta}=\ln\left[\frac{(N_{GG}+0.5)\times(N_{BB}+0.5)}{(N_{BG}+0.5)\times(N_{GB}+0.5)}\right]$ 近似于常态分布。

$$N\left(\mu = 0, \sigma = \left(\frac{1}{N_{GG}+0.5}+\frac{1}{N_{GB}+0.5}+\frac{1}{N_{BG}+0.5}+\frac{1}{N_{BB}+0.5}\right)^{\frac{1}{2}}\right)$$

步骤 3：检定规则。

- $\ln\hat{\theta}\sim 0$，表示晶圆图上故障晶粒或正常晶粒呈现随机分布状态。
- $\ln\hat{\theta}>0$，表示晶圆图上故障晶粒或正常晶粒呈现群聚分布状态。
- $\ln\hat{\theta}<0$，表示晶圆图上故障晶粒或正常晶粒呈现离散分布状态。

表 12.4 是其中四种不同类型的晶圆图，每片晶圆共有 268 颗晶粒。观察各片 $\ln\hat{\theta}$ 值，发现 No.1 与 No.2 的 $\ln\hat{\theta}$ 值趋近于 0，其图形的表现亦无法拒绝 H_0，即晶圆图上故障晶粒

或正常晶粒呈现随机分布。而 No. 3 的 $\ln\hat{\theta}$ 值为偏高的正值,表示图形有明显的群聚特征,同理 No. 4 的 $\ln\hat{\theta}$ 值为偏低的负值,表示图形有高度离散特征,发现该晶圆图有光罩重复相同错误(mask repeat error)的情况。

表 12.4 四种不同类型的 $\ln\hat{\theta}$ 统计报表

编 号	No. 1	No. 2	No. 3	No. 4
不同类型晶圆图				
$\ln\hat{\theta}$ 值	**0.102**	**0.006**	**2.876**	**−2.764**
N_{GG}	875	19	424	806
N_{BB}	2	712	357	0
N_{GB}	46	116	94	81
N_{BG}	42	118	90	78

根据检定规则的选取,可将图形群聚、随机或离散的情况进行分类(Taam & Hamada,1993)。根据 $\ln\hat{\theta}$ 检定结果,先将 138 片晶圆图分为四大类:

- **第一类型:随机性的晶圆图**。在此选取标准常态累积概率值介于 0.4 ~ 0.6 之间的晶圆图为随机性故障的图形,共挑出 9 片,其中又可根据芯片故障比例分成低度、中度、高度三种不同类别,如图 12.11 及图 12.12 所示。

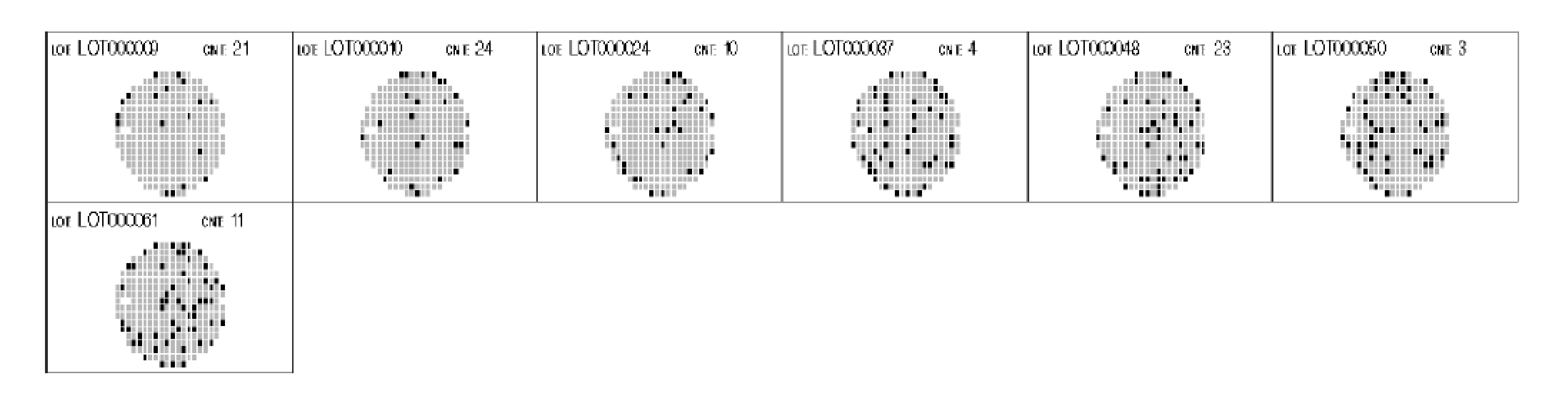


图 12.11 随机分布的晶圆图且故障比例轻微(CNT 表示此批货所含晶圆片数)

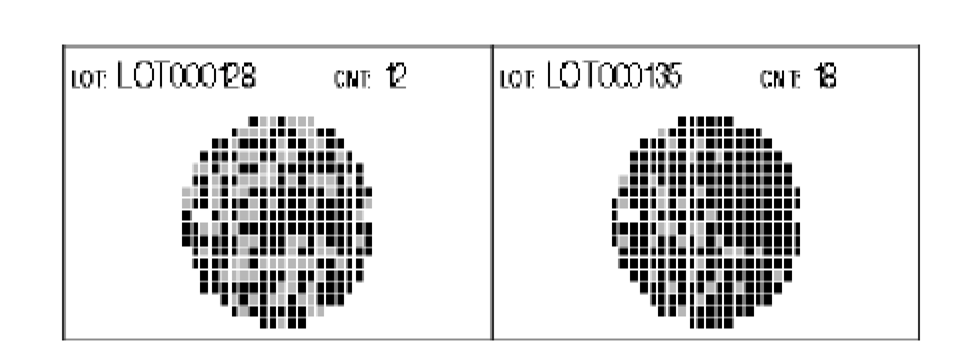


图 12.12 随机分布的晶圆图且故障比例严重

- **第二类型:群聚现象的晶圆图**。检定统计量出现较高正数者,可挑出 53 片具有显著

的群聚结块现象的晶圆图进行 ART1 分群。

- **第三类型：光罩错误的晶圆图**。检定结果显示 138 片晶圆图的 $\ln\hat{\theta}$ 并无较大的负数值产生，亦即无光罩错误的情况出现。
- **第四类型**：其他剩下晶圆图共有 76 片，归为第 4 类型等待进一步分类。

3. 强化特征，过滤噪声

本研究利用进退化法则，减少晶圆图上的噪声并强化特征样型，以显现晶圆图中特殊明显的样型，而加强图形代表性，本研究考虑正常与故障晶粒在二维空间上的图形，针对晶圆上某一晶粒周围相邻的 8 个位置给予权重，若上下左右的位置故障则令其权重为 1，斜对角的位置故障则令其权重为 0.5，若没有故障发生则其权重为 0，如图 12.13 所示。原始数据中，若某良好芯片位置周围的 8 个方格总和权重值大于或等于 4 时，则此位置予以进化改视为故障晶粒，如图 12.13(a)所示。过滤噪声的目的是希望滤除一些随机性的单一故障晶粒，减少图形比对上不必要的干扰，原始数据中，若某故障晶粒位置周围的 8 个方格总和权重值小于 1 时，则此位置予以退化，如图 12.13(b)所示。

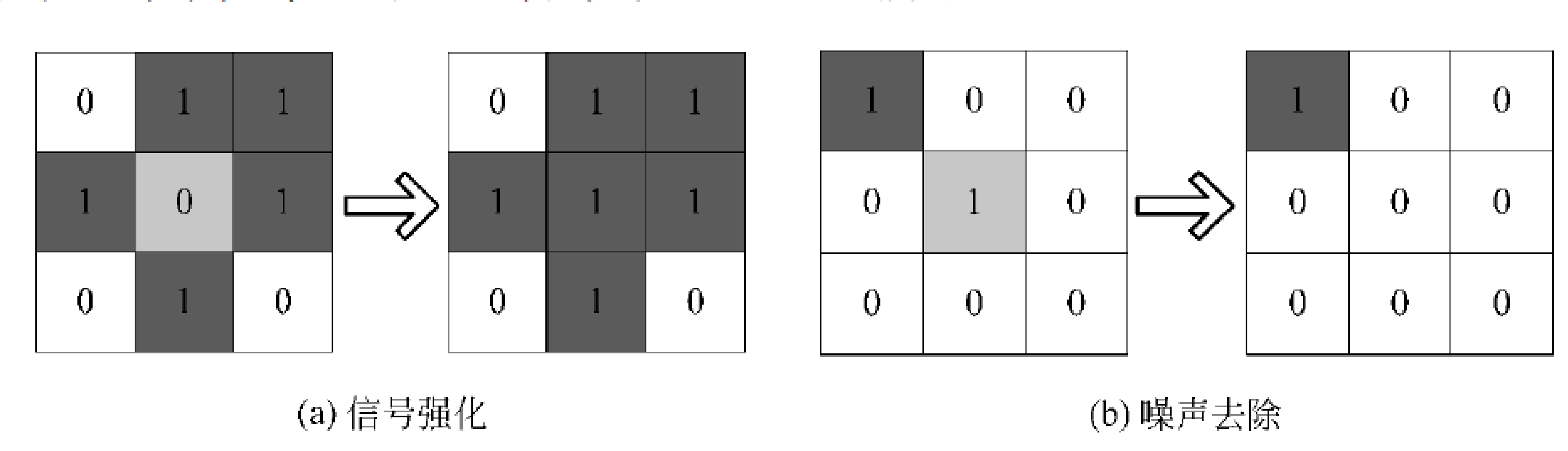


图 12.13　进退化原则示例(深灰格子表示故障晶粒，白格子表示非故障晶粒)

4. ART1 晶圆图分类

本案例中以每一群整理其故障晶粒位置的交集，成为共同故障特征(common failure pattern)，供工程师不需再逐一比对过去发生问题的所有晶圆图，达到缩小事故诊断范围的功效，之后可再进一步从具有相同特征的晶圆图中，去寻找过去也曾同样发生异常问题，而导致相同故障样型的原因，提供给工程师更多的诊断线索，作为发展晶圆图知识管理系统的基础。

已归属为随机性故障的晶圆图，不再予以分群。对于具有特殊样型的晶圆图则进行 ART1 分群，产生分群后的共同故障特征样版，此时再将原本暂时归入第四类的晶圆图与这些样型进行相似性比对。相似度高者则归属于同一群；相似度不高者，则再进行一次 ART 分群，以产生最后结果。

首先将归类为有群聚现象的 53 片晶圆图，先以进退化强化晶圆图中群聚的特征，再进行 ART1 图形分群，在警戒门槛值 $\rho=0.6$ 的情况下，可将 53 片有群聚现象的晶圆图分成 24 群，因此产生 24 个共同故障特征的样板。

同样地，若将其他类型的 76 片晶圆图，先进行进退化法则，再与这 24 个样板进行相似性比对，若相似度高，则将其归属样板中的某一群，若其相似度不高，则再进行 AR1 算法以得到分群结果。对于相似性的判断对标准，选择相似度超过 0.7 的值以表示有较高的相似度。在 76 片晶圆图中，有 15 片可以归属于原先 24 个分群的某些群组中。至于剩下的 61

片，则全部再作一次 ART1 的图形分类，并且选择较低的门槛值，群数才不至于过多，而失去分群的意义。因此选择相似度门槛值为 0.4，最后产生 47 群。

5. 结果诠释与评估

针对上述分类结果，请个案半导体厂内有一至七年不等晶圆图分类经验的十位领域专家，协助确认以此判断法分析评估晶圆图分类结果和内容的定性特性来分析本研究方法的适切程度，以诠释结果并检验效度。

领域专家先以一般常用的目视方式分别对同一组晶圆图数据进行分群及分类判断，并与本研究的分析结果进行比较，以评估本方法的内容效度。研究发现这些专家在分群及分类的结果也呈现很大的差异，分群数从 7 群到 19 群，判别出有群聚的批数也从 38 批到 56 批不等，而这些结果与工作年资也无正向关系；然而针对某些特殊的系统化故障样型，特别是噪声少与故障严重时，则有相当一致的结果。换言之，从柏拉图分析的观点或依据“80—20 原则”而言，可有效先找出故障严重的特殊系统化故障样型进行分析。

另一方面，领域专家在分析过程所使用的时间也有极大差距，即使最快的一位专家也花了近 40 分钟分出 13 群及 44 批货有群聚现象，其他专家有的花了数小时才分出结果，而本系统只要 20 分钟即完成所有动作。可见人为主观因素及对空间图形辨识能力的差异，往往会造成检测标准的不一致与人为偏差，因此，本案例的方法不仅可大幅降低人为因素，也由于自动化分类而使得工程师得以增加更多时间在缺陷诊断与良率提升上，提升半导体制造厂的质量管理。

12.3.3 案例小结

本研究发展晶圆图分类架构，一方面利用空间统计的方法，解决图形二维平面上相对位置的关联性，另一方面，结合 ART 人工神经网络图形辨识理论，处理数据转换后图形一维二元向量的相似性，因此同时考虑图形辨识上相对与绝对的观念。本研究可从大量的晶圆图数据库中挖掘出特殊样型和潜在有价值的信息，并将提取出的共同特征，构建某半导体厂的特征晶圆图图库，以发展结合晶圆图分类、领域知识及事故诊断的系统。由实证结果可以发现在分群好坏的群间距离要大，且群内距离要小两个主要指标中，ART1 人工神经网络其辨识效果在致力使群内变异最小的条件下能够获得最佳表现。

对半导体厂而言，可以有系统地将过去所产生的数以万计的晶圆图做分类，累积重要的制造智能，未来也将发展为在线实时晶圆图分类及比较功能，然后辅以领域专家的事故诊断经验，以提供工程师更多事故诊断线索，并通过将内隐知识外显化与系统化的过程，达到组织知识管理提升制造智能的目的（Chien *et al.*，2013；Liu & Chien，2013；Hsu & Chien，2007）。

12.4 低良率事故诊断与制程关联分析

12.4.1 案例说明

晶圆制造是以批量为加工单位。首先将晶圆激光刻号后，经过清洗（cleaning）送到热炉管内加热，在含氧的环境中，以氧化（oxidation）的方式在晶圆的表面形成一层二氧化硅

(SiO_2),紧接着以化学气相沉积(chemical vapor deposition,CVP)的方式将厚 1000～2000Å 的氮化硅(Si_3N_4)层沉积在刚刚长成的二氧化硅上,然后将在整个晶圆上进行微影的制程,先在晶圆上上一层光阻,再将光罩上的图案曝光到光阻上面产生显影。接着利用化学蚀刻(etching)或电浆蚀刻的方式,除去未被光阻保护的部分氮化硅层,留下所需要的线路图。再以磷为离子源,对整片晶圆进行磷原子植入(ion implantation),然后再去除光阻剂(简祯富等,2005)。

因此可依光罩所提供的设计图案,依次在晶圆上完成集成电路所需的晶体管及线路。接着进行金属化制程,制作金属导线,以便连接各个晶体管与组件,在每一道步骤加工完后都必须进行一些电性或物理特性的监控或测量,以检验加工结果是否符合规格;重复以上步骤一层一层地在硅晶圆上制造晶体管等电子组件,如图 12.14 所示。

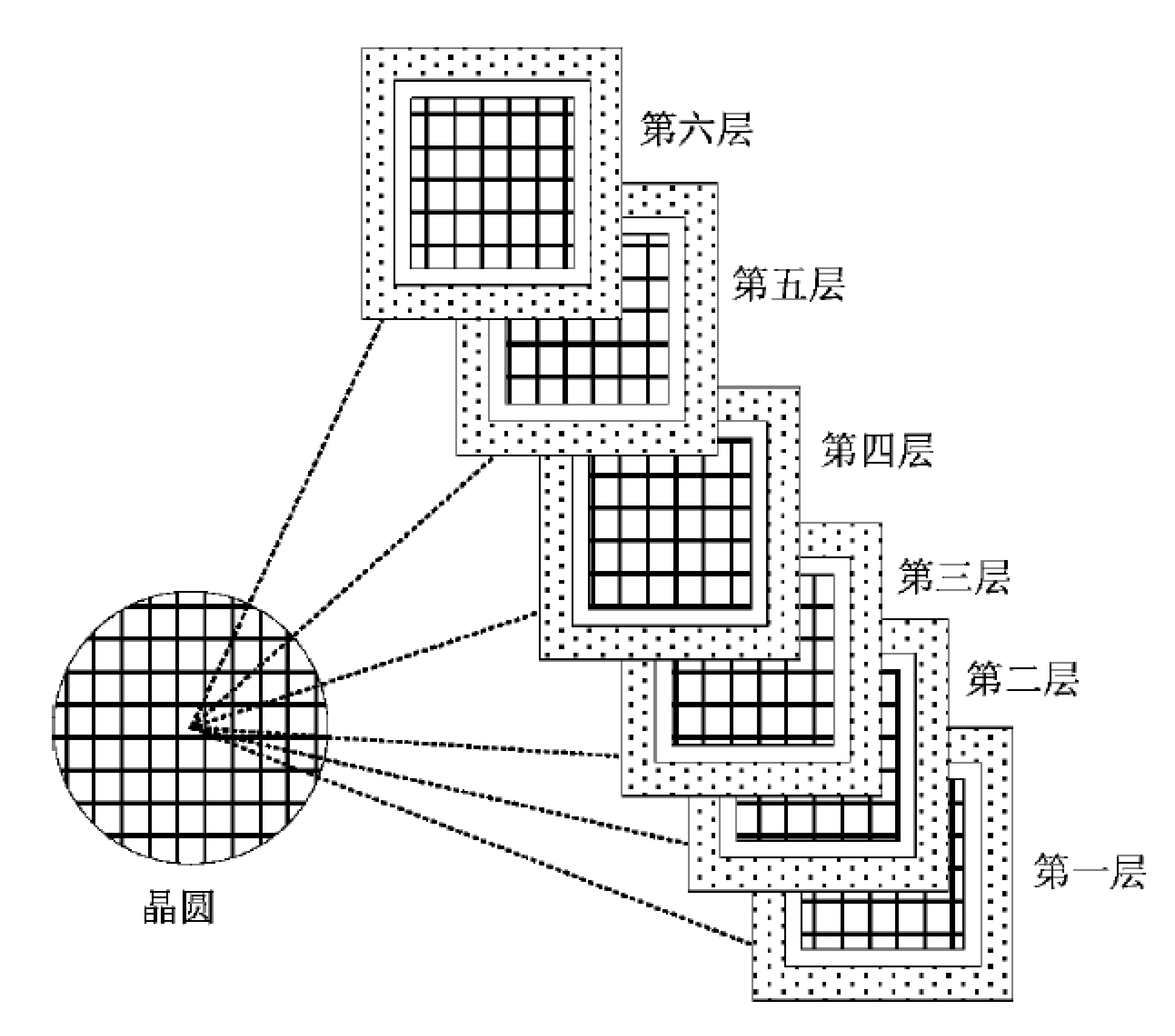


图 12.14　晶圆加工示意图

完成半导体制造流程后,晶圆允收测试与晶圆针测为确保生产晶圆质量的检验测试。晶圆针测检验主要目的在于确保制造后的每粒芯片的功能函数相符客户给予的规格。本案例即针对未通过晶圆针测的异常产品,从生产数据中试图找出有可能相关的制造流程站点、机台或使用配方设定等。当产品面临良率偏低的情况时,工程师需尽快找出可能的原因并修正异常。

然而,造成晶圆低良率的原因往往复杂且难以仅由单一个别原因完全解释,本案例(Chien *et al.*, 2007)建立一整合数据挖掘架构以分析生产数据与低良率间的关系,借由挖掘结果的累积,将其转换成系统性的规则或提取成知识,以供后续类似问题发生时得以有效且快速解决,并建立工程数据分析系统(Peng & Chien, 2003),当制程发生异常时,系统可自动产生信息提醒工程师注意,降低制程或机台异常带来的良率损失。

12.4.2　分析过程

1. 数据准备

在半导体前段制程过程中,不仅会记录过站制程与机台名称,也会记录过站时间的日

期、时、分、秒，若将时间当作变量进行分析，可能会因为过细的数据分辨率使得数据分析结果不佳，因考虑到日期的因素，所以需对日期数据进行转换，将机台及日期变量通通结合成一个新变量，即使过站机台相同，但只要过站时间不同，仍视为不同的数据。

本案例由工程数据数据库中，搜集某年度自 7 月 2 日至 8 月 20 日将近 77 批晶圆生产的生产数据与晶圆良率数据，其良率趋势图如图 12.15 所示。工程师欲从该检测结果回溯厘清制程发生异常的站别或机台，以尽快改善缺失。经由数据准备删除不需要的变量字段与离群值，且修正数据不一致后，最后从工程数据库中整理出共 71 批晶圆供后续分析。每个观测值包含晶圆检测良率数值、晶圆检测时间、作业阶层数目、作业机台名称、作业时间、晶圆批标识符等，共包含 455 个站的机台数。

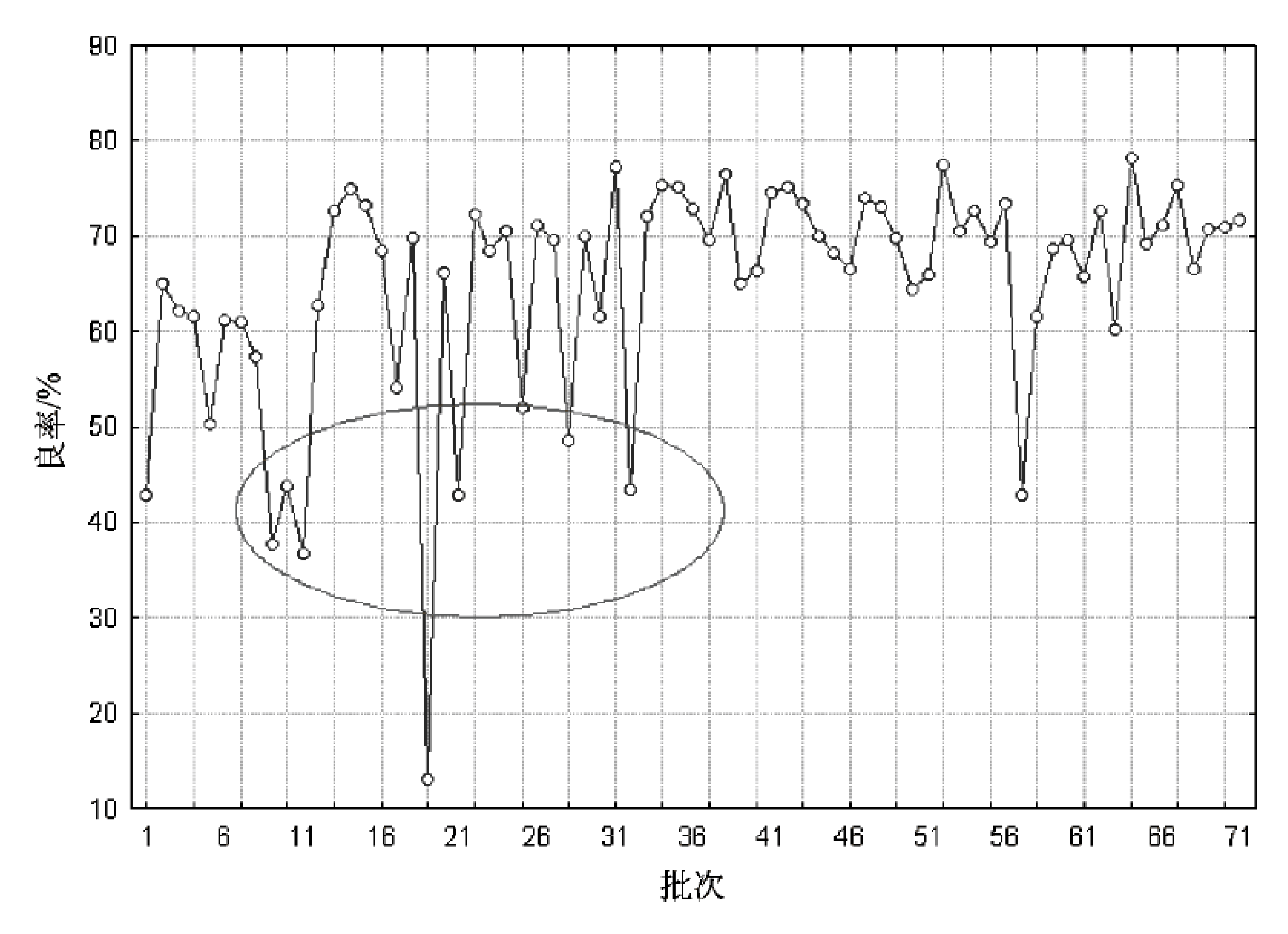


图 12.15 良率趋势图

首先利用聚类分析将数据转换为高低良率两群，接着筛选出显著影响低良率的异常制造站别或机台，以提高后续事故诊断分析效率。

以 K 平均法将 71 批晶圆分成高低良率两个群组，门槛值为 57.4%，其中有 59 批晶圆分至高良率群组。另外 12 批则分至于低良率群组，此两群组的基本统计量汇整如表 12.5 所示。

表 12.5 两群组的基本统计量汇整

群　组	批次数目	平均良率/%	良率方差/%
1（高良率批次）	59	69.491	23.740
2（低良率批次）	12	42.367	113.613

2. 数据挖掘模式构建

为了找出可能的异常机台，本研究利用 K-W 检定法（Kruskal-Wallis test）以检验某个制程下不同机台间的良率表现是否一致，若检定结果 p-value 小于显著水平，则表示不同机

台间具有显著差异,代表经过该制程的机台可能会造成不同的良率表现。

例如,某 M 站别中,有 3 台不同的机器设备,12 笔参数 A 的测量值如表 12.6 所示,以下为 K-W 检定的执行步骤。

表 12.6　某 M 站中 3 个机台的参数 A 的测量数据

测量参数 \ 机台	E-1	E-2	E-3
A	9.10	9.06	9.27
	9.41	9.00	9.15
	9.07	9.01	8.98
	9.03	8.72	8.91

步骤 1:检定假设。

H_0:M 站中的 3 台不同机台,其测量参数值 A 的表现皆无差异。

H_1:M 站中的 3 台不同机台,至少有一台测量参数值 A 的表现有差异。

步骤 2:检定统计量。将观测值依递增顺序列出各 R_{ij},如表 12.7 所示,因此根据式(12.6)与式(12.7),可算出检定统计量 $H=7.423$。

$$H=\frac{1}{S^2}\left[\sum_{i=1}^{k}\frac{R_{i.}^2}{n_i}-\frac{N(N+1)^2}{4}\right] \tag{12.6}$$

$$S^2=\frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{N_i}R_{ij}^2-\frac{N(N+1)^2}{4}\right] \tag{12.7}$$

表 12.7　原始数据的秩排序

参数值 \ 机台	E-1	E-2	E-3
A	9.10(9)	8.91(2)	9.27(11)
	9.88(12)	9.00(4)	9.15(10)
	9.09(8)	8.98(3)	9.01(5)
	9.03(6)	8.52(1)	9.06(7)
$R_{i.}$	35	10	33

步骤 3:检定规则。在置信度 $\alpha=0.05$ 的情况下 $H=7.423>\chi^2_{0.05}(2)=5.991$,故拒绝(reject)M 站中 3 台不同的机台,其测量参数值 A 的表现皆无差异的假设。

步骤 4:结论。借由观测统计报表中的 p-value 字段亦可做出相同结论,其 p-value 越小,表示有越足够的证据显示 M 站中 3 种类型不同的机台在参数值 A 的表现上确实有差异。

步骤 5:诊断。将所有的站别,重复执行步骤 1~4,最后算出所有 p-value 后,由小到大作排序,即可找出发生故障概率较高的前几站。

半导体制程的机台数目众多,本案例中变量个数远大于样本数据个数,利用无母数 K-W 检定找出显著影响良率的制程站点与机台,针对晶圆良率分别检定以下之虚无假设以及

对立假设：

H_0：某站别中的 n 种类型不同机台，其良率的值无差异。

H_1：某站别中的 n 种类型不同机台，其良率的值有差异。

共 455 个站分别进行 K-W 检定，以计算出检定的 p-value，再由小至大排列如表 12.8 所示，并与工程师讨论与设定定义显著影响的站点门槛值为 0.3，经过筛选后剩余 168 个站点被视为候选站点。

表 12.8　K-W 检定结果(部分节录)

顺序	站别	p 值	顺序	站别	p 值	顺序	站别	p 值
1	182	0.000	11	95	0.010	21	115	0.025
2	2	0.001	12	119	0.011	22	7	0.026
3	41	0.001	13	54	0.012	23	359	0.026
4	192	0.004	14	436	0.012	24	397	0.028
5	93	0.004	15	163	0.015	25	172	0.030
6	210	0.006	16	225	0.017	26	75	0.032
7	94	0.007	17	52	0.017	27	183	0.032
8	103	0.009	18	20	0.018	28	208	0.032
9	124	0.009	19	64	0.022	29	230	0.032
10	170	0.010	20	42	0.024	30	252	0.032

注：因篇幅关系，在此仅列出 p-value 由小至大排列的前 30 个站别。

本研究先以高低良率聚类为决策树目标变量，再将每一批晶圆过站所使用的机台和时间当成变量进行决策树分支。在诊断半导体制程异常时，有可能在决策树的第一层就可以解释大部分发生异常的原因，然而，却有些少部分的原因是来自于第二层或是更往下的层次才能提升此规则的解释能力。因此可以借由决策树的分支，察觉出一些工程师不易从本身专业知识得出的信息，或不容易由第一阶层就找出显著的异常发生原因。

本实证研究以 168 个关键站别所包含的机台数目以及流程时间为候选分支属性，并以良率数值为目标值，进而以 F 检定统计量找出第 261 站别为分支属性能表示最显著的分类结果，如图 12.16 所示。其中所有 71 个批货的平均良率为 64.906%，以站别 261 分支后，可发现在 6/13、6/16、6/26 以及 6/27 过该站别的机台一为造成产品异常的最主要原因，其平均良率百分比为 45.882%，因而作为决策树分支的依据。

3. 结果诠释与评估

工程师可经由数据挖掘的分析结果与规则，配合本身专业知识判断，快速找到造成低良率的可能原因，本案例应用数据挖掘技术，缩短可能的事故原因范围，如图 12.17 显示第 261 站别的机台 A 的良率表现。因此利用数据挖掘所挖掘出规则“自 6/13 后经过第 261 站别的机台一会导致异常品产生的概率相当高”。

本案例先找出关键站别与机台，进而提取显著分类规则以得知各属性的关联性，在分析

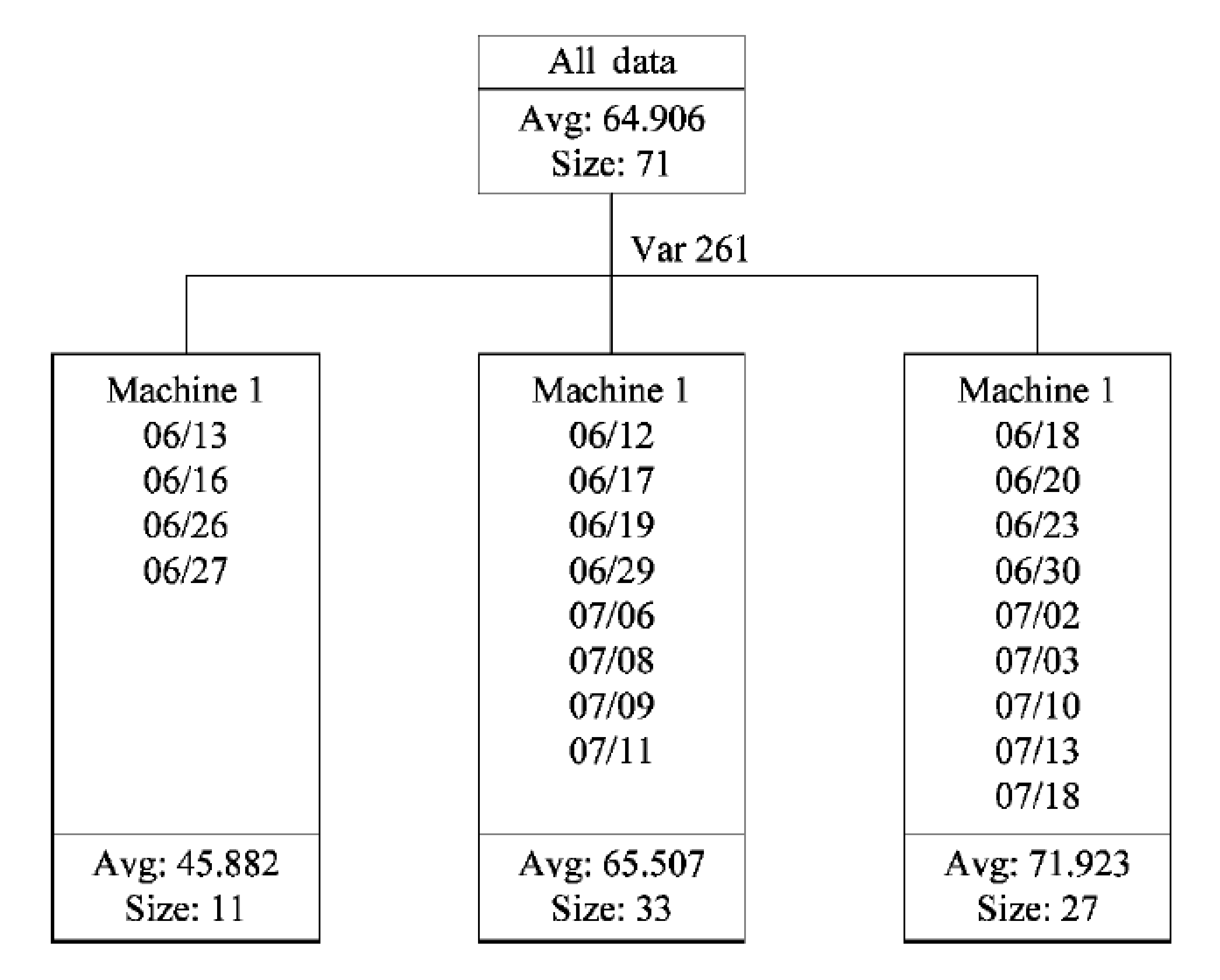


图 12.16　决策树分类结果

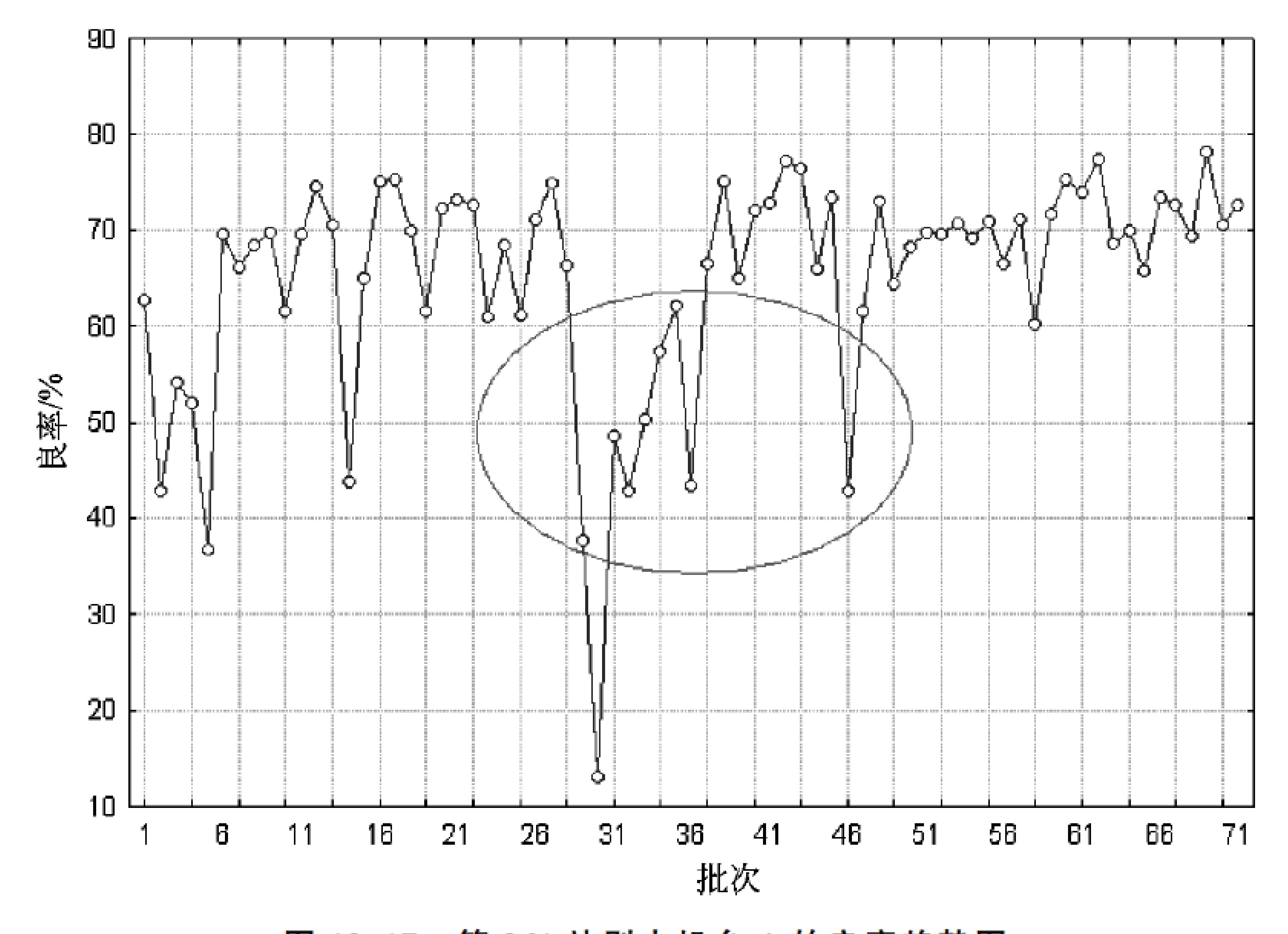


图 12.17　第 261 站别中机台 A 的良率趋势图

过程乃至于最后数据挖掘的结果，不论是数据、可视化图形或规则化叙述，应不断与领域专家讨论以获得其经验与进一步改良的意见。挖掘的结果对于工程师是否有帮助或整个挖掘过程是否达到预期效果，皆须通过结果解释与讨论重复循环，以厘清关键机台定义与规则提取所代表的意义与价值，才可使得研究模式与结果更加完备。

12.4.3　案例小结

使用 K-W 检定法验证在同一站中各机台的产出质量间是否存在显著差异性，需要注意的是某些站别仅有一机台，因此无法与其他站别进行标杆验证。此时，可收取该机台的过去

时间间隔的制程表现数据，以各区间的质量良率指数作为与该切断时间点的比较基准，同样以K-W检定方法验证该时间点的产出表现是否有显著偏异。工程师可根据K-W检定方法所得的检定值(常以 *p*-value 表达)，找出发生故障性较高的站别，加上专业知识的判断，以快速进行站别中机台事故诊断与紧急换线处理。

12.5 半导体制造管理的数据挖掘

12.5.1 案例背景

运用数据挖掘和大数据分析方法，从晶圆生产制程中累积的大量原始数据中提取特定的样型，可以得到一些实用的信息作为降低生产周期时间(cycle time)和制造管理决策的依据(Chien *et al*.，2012；Kuo *et al*.，2011；简祯富等，2004)。

本研究以目标层级架构方法推导半导体厂制造管理的绩效指标，发现到各个指标之间的连动关系与相互影响，例如生产线在制品数量(work-in-process，WIP)与机台利用率有正向关联，但却又必须和产品周期权衡。但是根据 Little's Law，产出量、生产周期时间、WIP与成本之间，也具有相关性，例如，当机器效率增加时，会有较低生产周期时间和较低的在制品存货；当在制品增加时，则生产周期时间也会增加。

12.5.2 分析过程

1. 数据准备

本案例以某半导体公司提供的35万笔生产数据，说明如何应用数据挖掘在制造管理上。个案公司为自有品牌的半导体厂，月产量约4万片晶圆，基本的产品组合可分为两大类：标准型产品以及接单生产产品，而接单生产产品又有约2/3的制程可以事前计划生产，其半成品则储存在线，在数据的处理上不将此半成品列入WIP计算，因为这类半成品将停留一段不等的时间，待接到订单后，才由半成品库送至生产线继续剩余制程到出货。

本案例搜集将近9个月的半导体生产数据，数据形态为每日各个机台生产状况。包括所有的生产流程(route)以及各产品制造Layer与关键指标，例如每日生产量(daily move)、WIP、设备使用率(utilization)、转换率(turn ratio，TR)、产品组合(product mix)等。

WIP与T/R的数据分布如图12.18和图12.19所示，其中，图12.18中空心方形的T/R和WIP关系呈现上升的趋势，表示在该段时间内产能为扩增的状况，造成WIP越高，T/R也越高。另外，在图12.19中可得知投片量有很大的差异外，在产品组合也因时间不同而变化。为避免数据变异过大，影响对数据的长期趋势分析，所以在数据分析前将数据作4期的移动平均。

2. 产品组合分群与数据转换

依据不同的产品组合与WIP水平利用SOM找到产品组合与WIP间的群聚关系，其结果如图12.20，共可分为五群，各聚类与产品组合的比例结果如表12.9。从图12.20拓扑图中，发现不同群的数据中，哪些产品是该群的主要产品。若综合聚类结果、WIP与生产日期，可以明显知道大部分数据与时间均有相关，也跟实际的投片计划相符合，如图12.21。

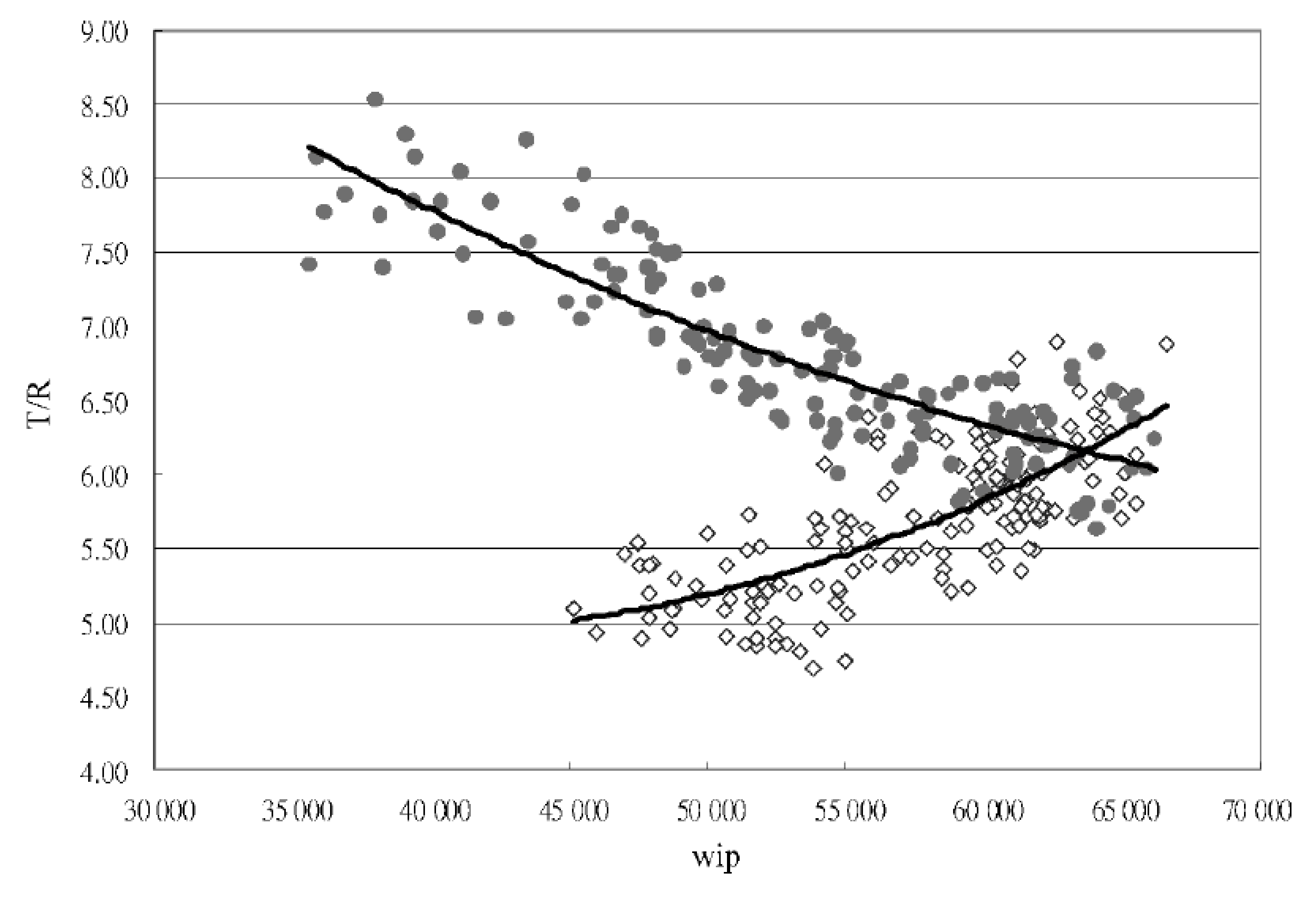


图 12.18　WIP 与 TR 的关系图(简祯富等,2004)

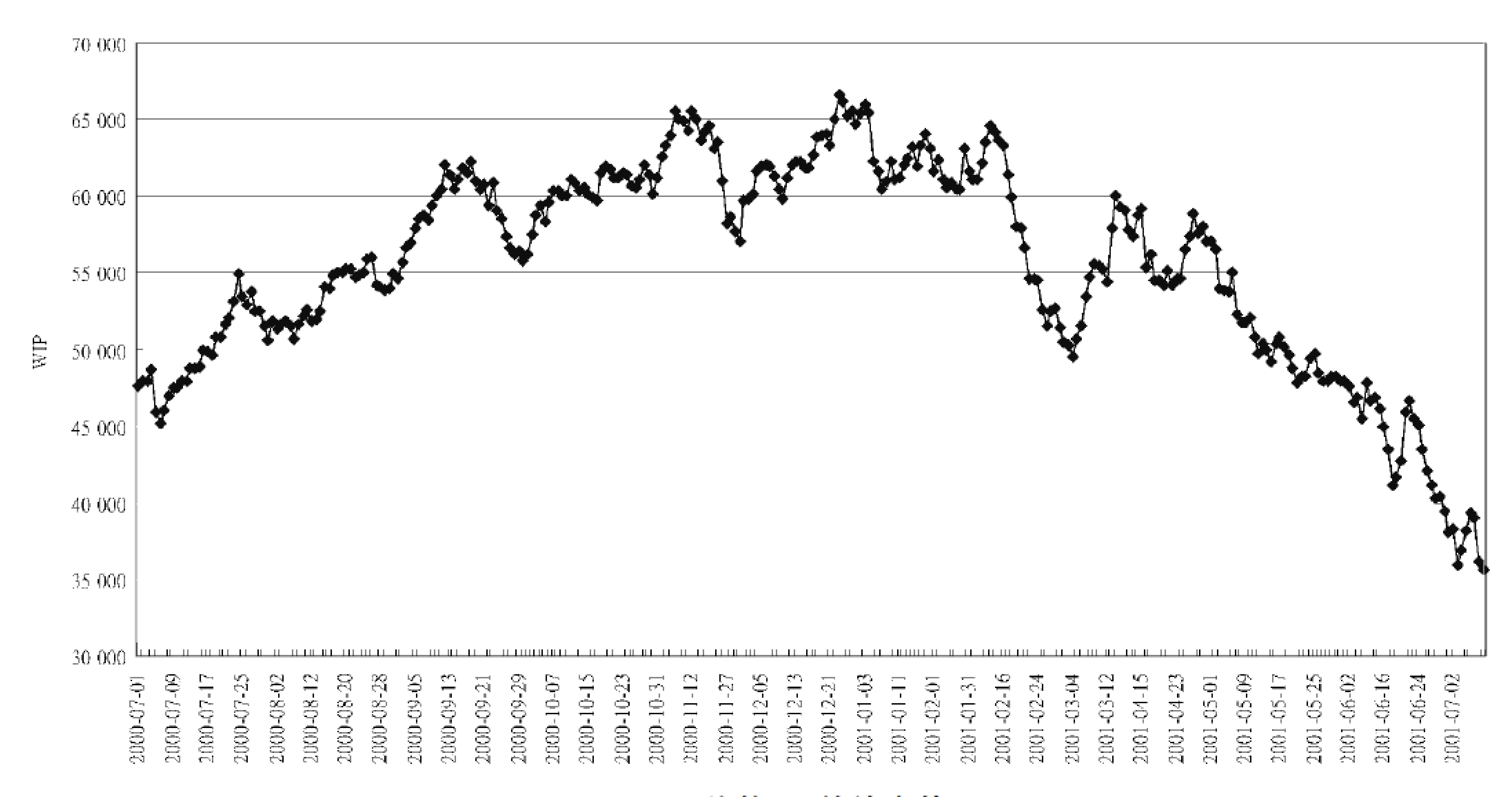


图 12.19　WIP 趋势图(简祯富等,2004)

表 12.9　产品组合分群结果

聚 类	产 品 组 合								
	WIP	P1	P2	P3	P4	P5	P6	P7	P8
1	52 062.03	0	0.043	0.384	0.258	0.021	0.239	0.030	0.024
2	58 949.85	0.052	0	0.328	0.252	0.012	0.306	0.019	0.031
3	53 493.28	0.211	0	0.232	0.227	0.022	0.262	0.017	0.029
4	63 497.42	0.204	0	0.197	0.214	0.124	0.114	0.117	0.030
5	66 710.91	0.178	0	0.073	0.245	0.247	0.047	0.150	0.059

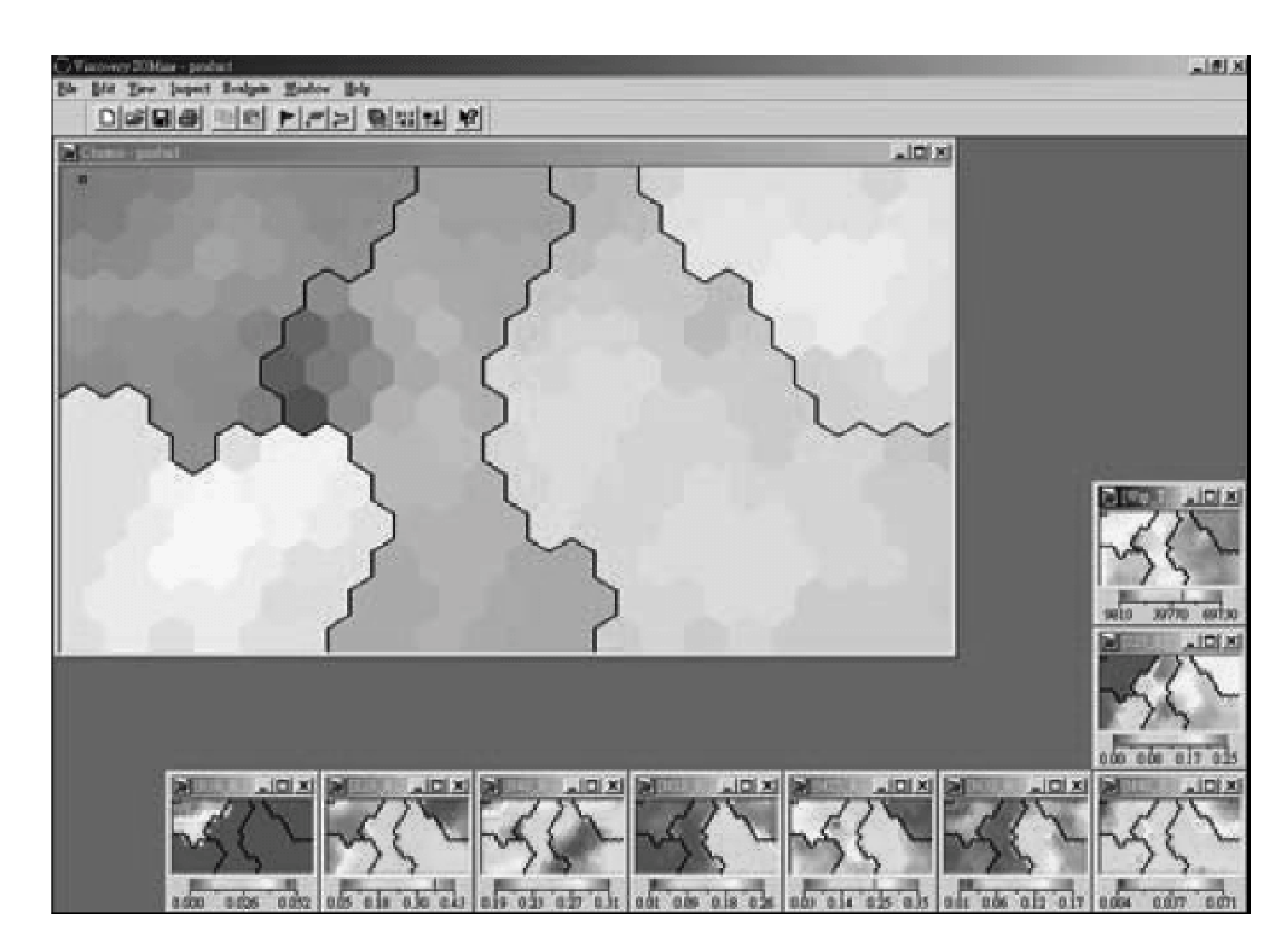


图 12.20　产品组合拓扑图群聚现(简祯富等,2004)

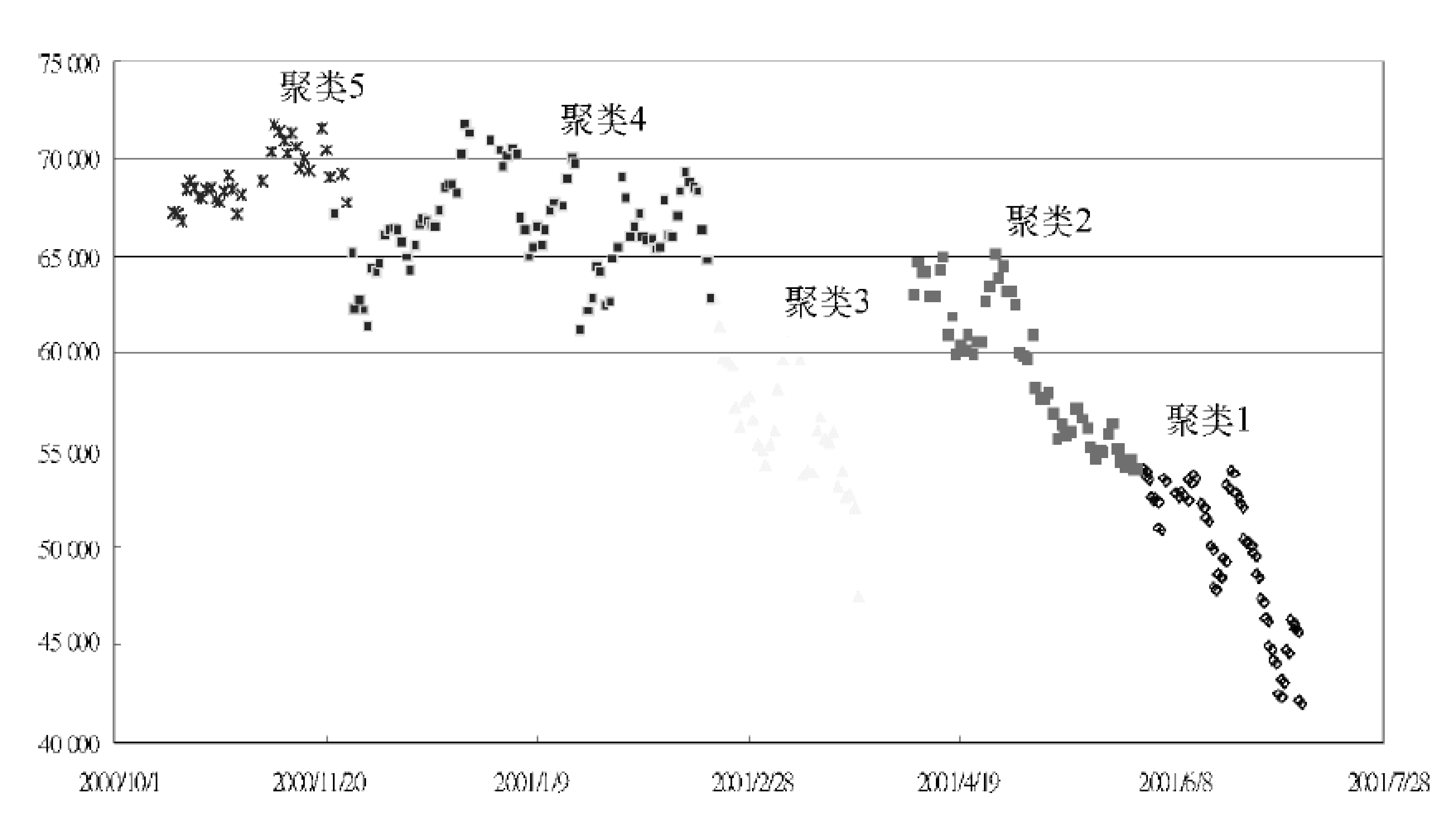


图 12.21　产品组合分群结果(简祯富等,2004)

3. 决策树分析与分类规则提取

借由与领域工程师的讨论后,可依据这些产品组合的改变引入拓扑图中,在不同的产品组合变化下提供不同的管理依据。

4. 决定 WIP 的水平

利用第 4 章决策树分析,以 Move 为目标变量,依分群以及领域知识定为低、中、高三个等级,如图 12.22 中的深色横线,再以 WIP 为分支变量,可将 WIP 划分为 WIP<62 500、62 500≤WIP<70 055、WIP≥70 055 三个水位,根据决策树规则结果与图 12.22,WIP 在大于 62 500 时,Move 可有较多的产出。

再加入 T/R 对 Move 的情况,经由决策树分析结果,可发现当 WIP>68 000 时,T/R 的表现有 72%是在较低的状况。因此,可归纳得到在其他生产条件固定下,最适的 WIP 水位在 62 500≤WIP<68 000(图 12.23)。

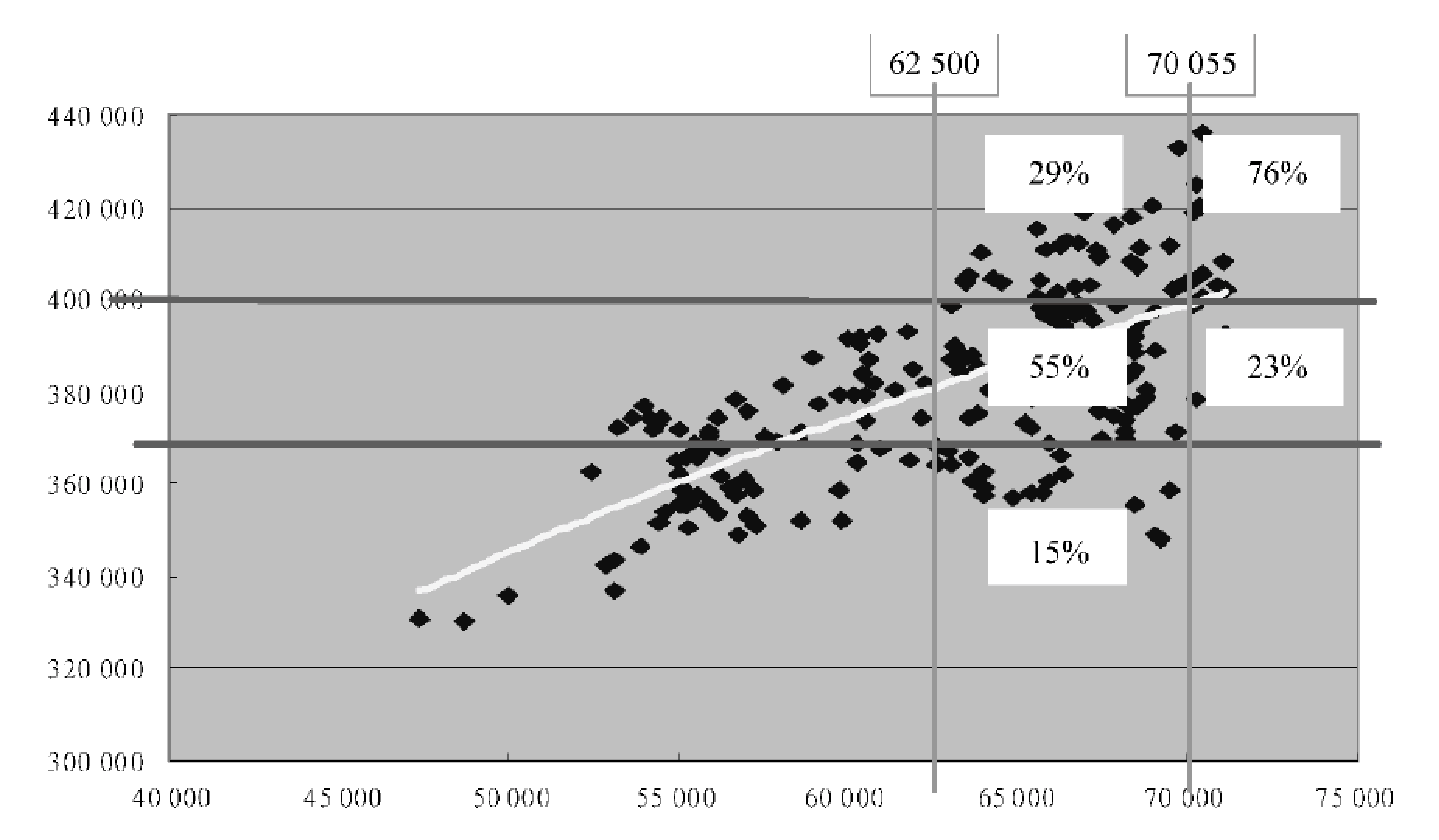


图 12.22　Move 对 WIP(简祯富等,2004)

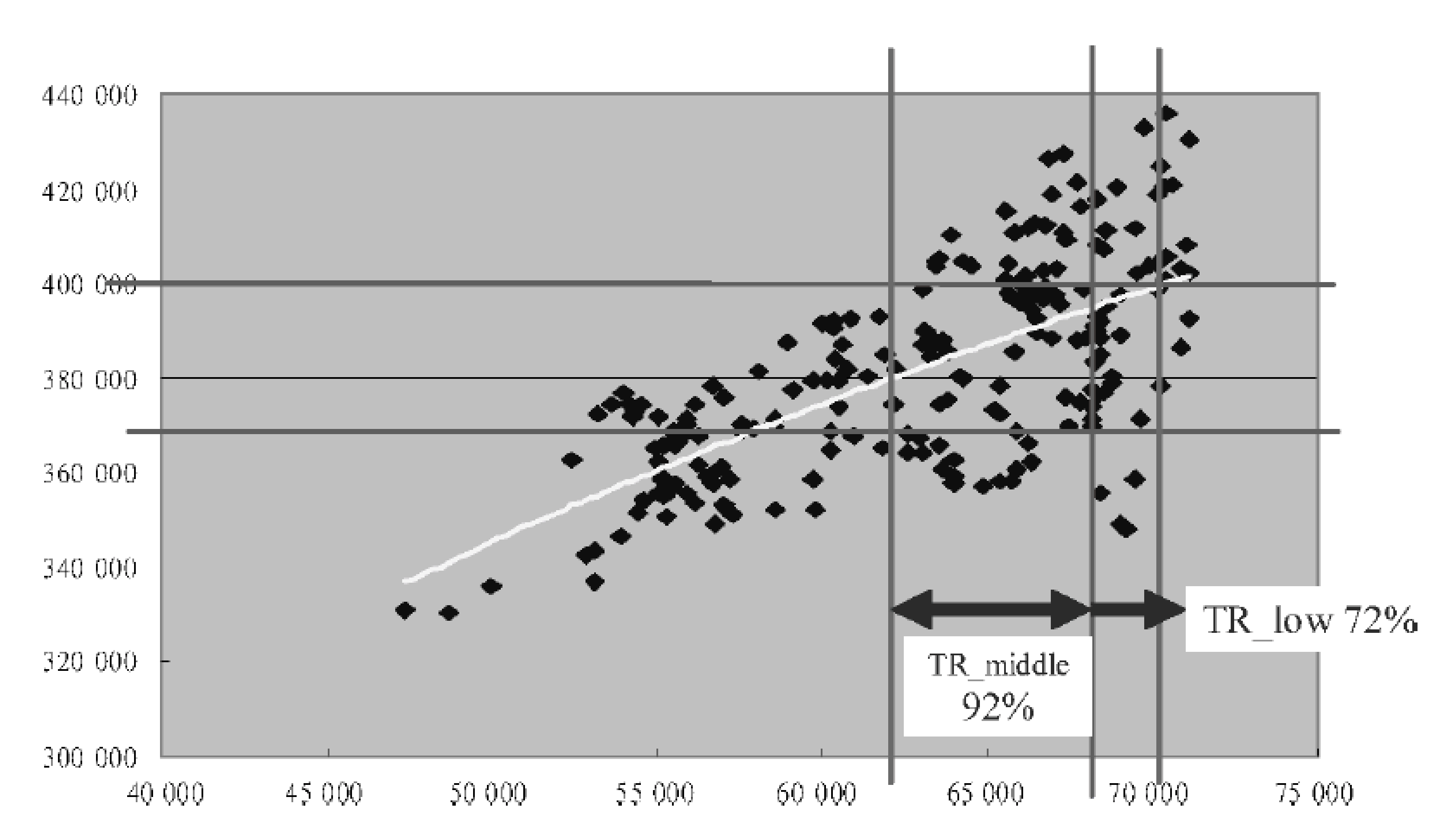


图 12.23　利用 TR 找出 WIP 对 Move 的最适状况(简祯富等,2004)

5. 决定较佳的产品组合

分析产品组合对 Move 的影响,发现 WIP 的区间在 62 500≤WIP<68 000 时,若接单生产产品占全部 WIP 的比率超过 0.48,此时 WIP 必须维持较高的水位,而比率在 0.385 以下在 Move 的表现均较比率 0.385 以上为佳。如图 12.24 所示,WIP 的区间在 62 500≤WIP<68 000,接单生产产品的比率在 0.385 以下,在 Move 的表现会比较好。制造现场管理者可以根据不同生产状况,调整生产线 WIP 的数量。

6. 结果诠释与评估

根据产品组合分群的结果,再加入机台使用率对 Move 的影响。以决策树 CHAID 分析在每个分群中影响到产出量的主要机台,以提供决策者在规划投片计划时,预先规划机台的状况,避免因规划不良造成产出损失。

例如,第 4 个聚类中共有 80 笔数据,其平均的 Move 量为 396 737,图 12.25 为聚类 4 利用决策树分析的结果,聚类 4 共得到 11 条规则如表 12.10。可得到当 I-Line 机台的使

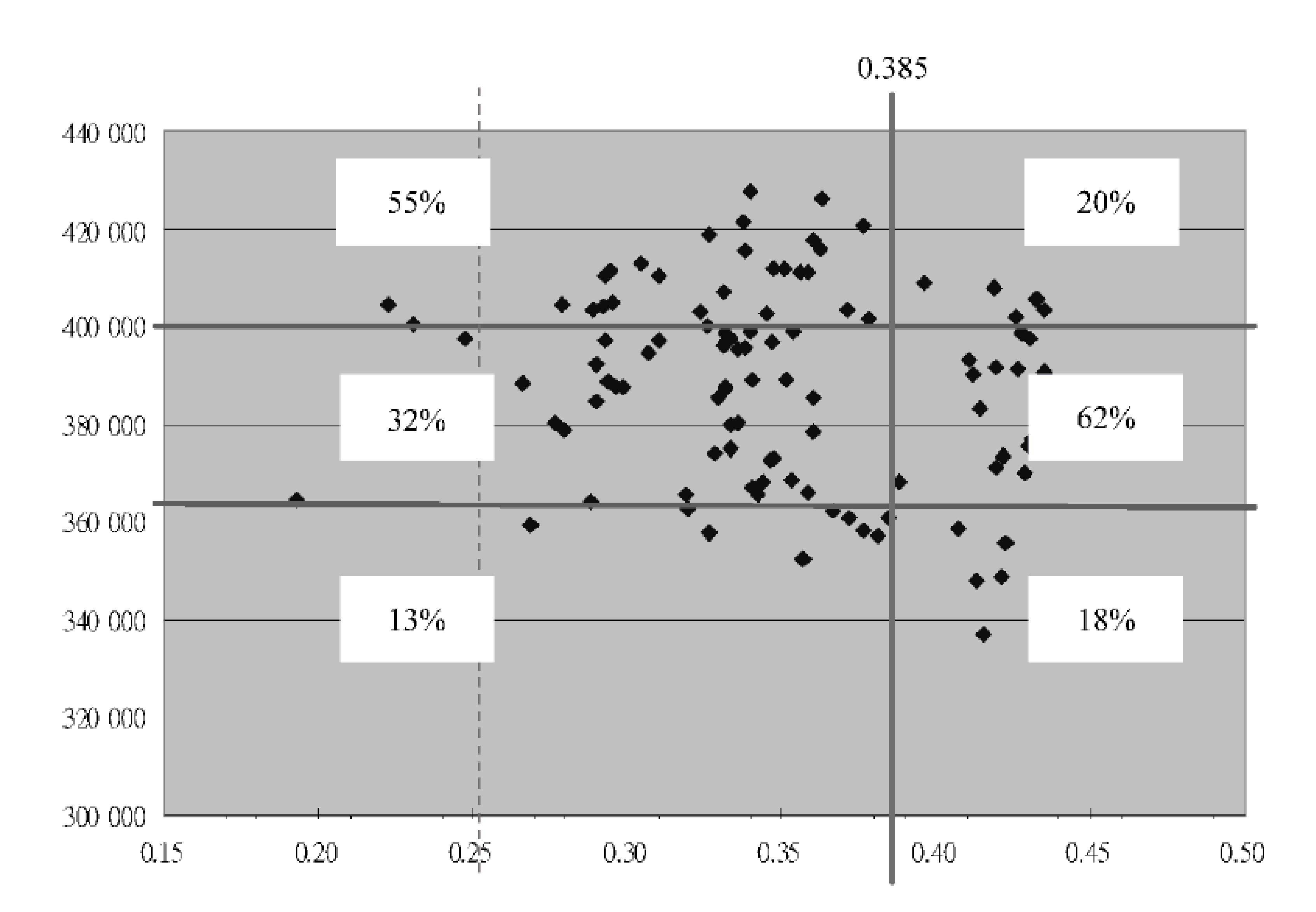


图 12.24 产品组合对 Move 的影响(简祯富等,2004)

用率大于 85%时,有 33 笔数据其平均 Move 量为 403 647;当 I-Line 机台的使用率小于 85%时,有 47 笔数据其平均 Move 量为 391 886;当 I-Line 机台使用率大于 85%,且 CLSF 机台的使用率大于 70%时,有 29 笔数据其平均 Move 量为 409 122。也可使用决策树发掘其他 4 个不同产品组合下的设备利用率与 Move 的规则。

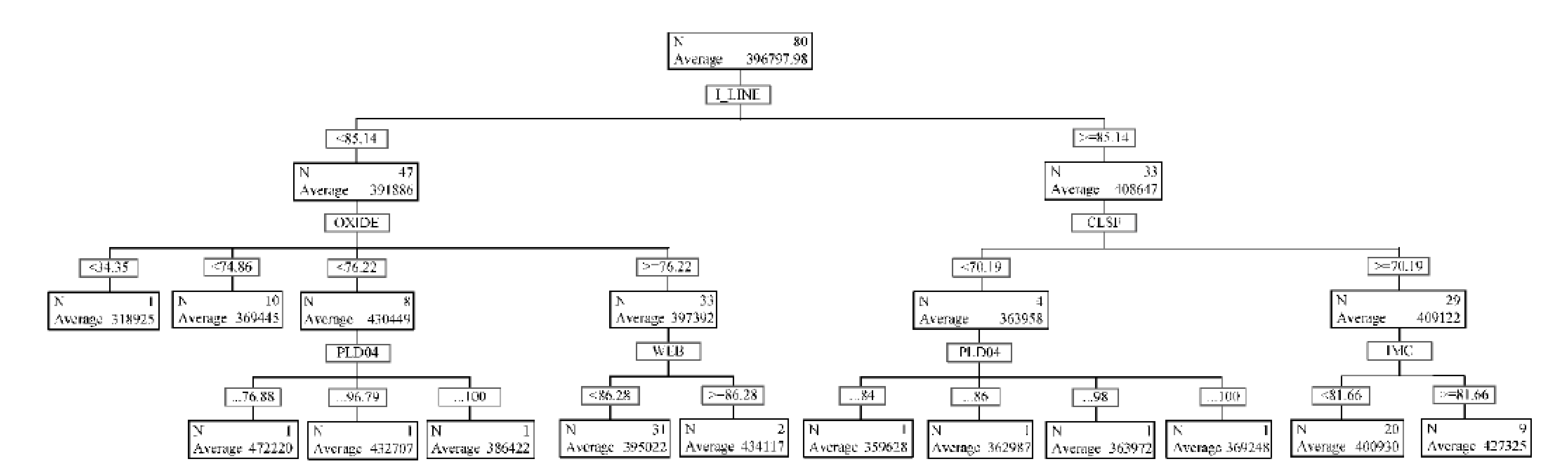


图 12.25 Cluster 4 产品组合下,机台使用率对系统产出的决策树分析图

表 12.10 Cluster 4 的决策树归纳规则

规　　则	数据量/笔	平均 Move 量
L-Line>85%	33	403 647
L-Line>85% & CLSF>70%	29	409 122
L-Line>85% & CLSF<70%	4	363 958
L-Line>85% & CLSF>70% & IMC>82%	9	427 825
L-Line>85% & CLSF>70% & IMC<82%	20	400 930
L-Line<85%	47	391 886

续表

规　　则	数据量/笔	平均 Move 量
L-Line＜85% & OXIDE＞76%	33	397 392
L-Line＜85% & OXIDE＞76% & WEB＞86%	3	434 117
L-Line＜85% & OXIDE＞76% & WEB＜86%	30	395 022
L-Line＜85% & 35%＜OXIDE＜74%	10	369 445
L-Line＜85% & OXIDE＜35%	1	318 925

12.5.3　案例小结

本案例从 WIP 与 Move 的实证分析中可知(图 12.26)，如果要确保较高的生产量，在制品数量大于 62 500 较容易获得高 Move 量。但是，加入 T/R 的分析，可得到区间 1(62 500≤WIP＜68 000)与区间 2(68 000≤WIP)，虽然两个区间在较高的 Move 差异有 12%，但是必须付出 T/R 较低的代价，本案例提供决策者权衡生产指标的方法。此外，产品组合在接单生产产品的比率上最好保持在 0.384。最后，提取在不同 WIP 水位下，设备使用率对 Move 影响，提供生产规划人员投片计划的参考，在产出与设备利用率间找到较佳的生产组合。

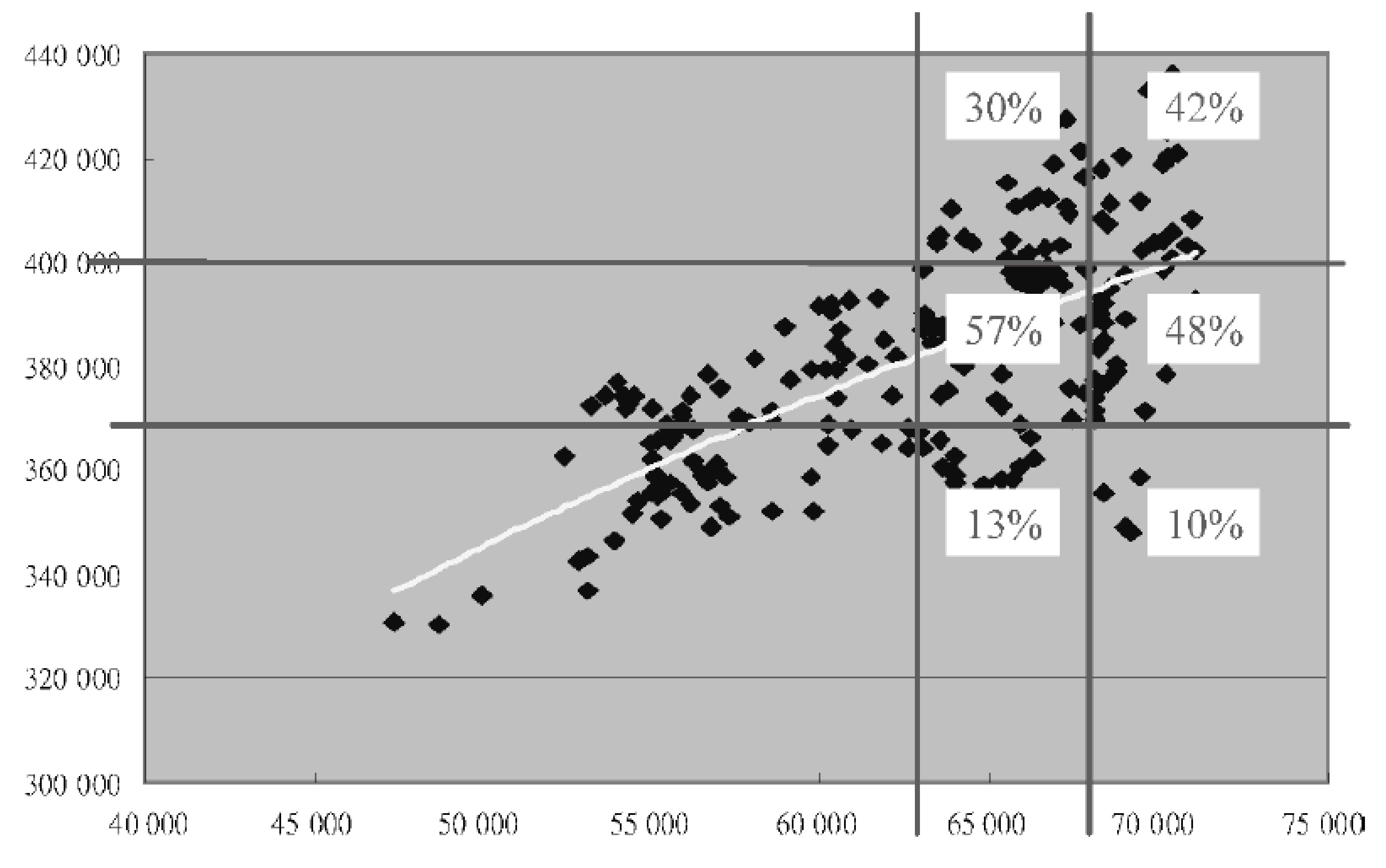


图 12.26　最适的 WIP 区间(简祯富等，2004)

台积电著名的知识管理，主要是通过各种技术委员会作为跨厂区单位的知识分享与标杆学习平台。然而，制造管理和其他结构化的制程技术不同的是，在某个厂区最佳的制造管理实务(best practice)并不一定能直接应用在其他地方，因此，须将复杂的实际问题架构成数学模式，并建立可以随时空环境转换的决策分析模式，并导入数据挖掘降低生产周期时间以提升生产力的方法。

半导体进入消费电子时代之后，产品价值随着时间快速折旧，因此上市时间和生产周期时间的缩短极为重要。另一方面，由于半导体的生产模式相当复杂，所以传统生产管理理论仅能处理小范围的工作站，Kuo 等(Kuo *et al.*，2011)利用半导体制造的巨量数据，分析影

响在制品水位和在线等候时间的影响因子，以找出每个工作站在线在制品的理想水位和产出关系，通过宏观调控机制以维持生产系统的平衡与加工流程的顺畅，有效地降低生产周期时间，并荣获美国电机电子工程师学会年度最佳论文（2011 Best Paper of IEEE Transactions on Automation Sciences & Engineering）。

12.6 结论

随着全球化的竞争及美国制造业复兴（US Manufacturing Renaissance），许多制造业面临着如何在提高产量与生产效率的同时保持和增加产品良率的问题。另一方面，2013 年 4 月，德国于汉诺威的工业博览会中，提出“工业 4.0”（Industry 4.0）的新兴概念，在智能制造逐渐成为新世代工业的核心之后，传统生产制造的商业模式、价值链、服务与分工形式将大幅改变，导向第四波的工业革命。因此有效运用大数据分析技术和制造智能的方法对生产制造过程进行有效的监控，以对已经出现的或将要出现的故障进行准确及时的诊断和排除，以提升良品质量和生产效能，已成为全球高科技产业的重要问题。

高科技制造产业的竞争优势取决于成本、质量以及达交时间，其中尤以质量为占有长期市场竞争优势的主因。数据挖掘的优点在于可发掘原始数据中所隐藏的有价值的信息，因此，若能借由系统化的分析从庞大的工程数据提取具代表性的信息并转换成有价值的知识，以提升高科技制造业产品良率、增加生产力、优化制造资源分配的决策辅助与知识参考（Chien & Hsu，2014；Chou *et al*.，2014）。

数据挖掘的目标可能是找出异常的参数，或是进行低良率产品之事故诊断或故障排除，因此需根据问题目标回溯（retrieve）相关的制程数据，选择适当的方法或模式进行挖掘，并不一定需预先设定问题的模式，而通常所得到的结果也往往是先前未知的。尽管每次事故发生的问题类型并非一成不变，仍然可依据系统性的数据挖掘架构，按部就班地进行分析。待累积足够的经验后，整理出系统化的规则和模式，以自动化方式进行例行性分析过滤有可能发生的问题，一旦发生特殊状况时，系统可立即呈现信息，进而达到系统化的最终目的。

台积电曾把晶圆厂自动化的发展分为拟人化、无人化、超人化三个阶段。也就是说，一开始是用计算机和设备学习人的做法，接下来是将机械性的工作自动化以取代人，最后则是发展一个集结众人智能的制造系统。让系统不仅能自动化，还能“智能化”地知道如何判断和决策，而超越一般人的能力。这不仅是未来趋势，也是极大的挑战（简祯富，2014a）。

半导体产品制程影响变量众多，存在复杂的交互作用，前制程参数常会影响后制程的良率，单靠专家知识判断不易解决。虽然数据挖掘可以有效率地从大量数据中提取有用的信息，但是若仅套用数据挖掘软件却不一定能够达到效果，特别是当数据本身有很多噪声和复杂的交互作用时。例如，利用主成分分析法来产生新的变量以降低解释变量相依性，然而却面临所产生的主因素的诠释问题，且对找出事故关键因素与后续的事故排除未必有用。

导入大数据分析以提升制造智能过程中，应持续与领域专家与工程师沟通讨论，不断地循环改善挖掘架构，可提取宝贵的信息和制造智能，协助工程师做出判断。因为半导体制程的数据挖掘过程中，很少只利用单一模式就可挖掘出所需的所有信息，而需针对事故问题不同的特性，使用不同的数据挖掘工具，不断厘清参数间的关联性来推论制程影响因素间的关联性，来逐步找寻原本受显着因子影响而掩盖其效应却可能真正影响产品良率的制程因素，

以有效提供事故诊断与排除的线索。而所累积分析结果与制程特性的关联性,也提供建立更完善的半导体数据挖掘架构,加速数据挖掘、信息提取与产业知识管理的系统化,以建立更完善的制造智能系统。

随着半导体产业进入大者恒大的竞争赛局,建造一座十二英寸厂至少投资四十亿美金以上,其中超过六成的资金都用于购买机台。半导体产业进入门槛高退出门槛也高,一次的投资还不够,在制程技术持续演进与产品的更迭下,必须每年更换或升级相关设备,才足以维持竞争力。然而,半导体产能建置扩充前置时间长,加上需求变动大、不确定性高等因素,都造成产能规划的困难,也影响客户需求满足和公司的成长与获利。

换句话说,半导体厂的产能规划决策往往必须在需求高度不确定下进行。因此,利用大数据分析技术,我们整合产品生命周期与技术扩散理论,检验产业环境的实际影响因子,发展考虑多世代技术扩散、技术替代、重复购买、价格、市场成长率和季节等因素的"产品生命周期和数据挖掘的需求估计技术",并建立一个可以随着时间推移而调整和更新需求预估模型的机制,作为预测未来需求以辅助制定中长程产能策略的依据;并结合最小化最大可能后悔(mini-max regret)的赛局策略(Chien &zheng, 2012),动态调整产能规划(Chien *et al.*, 2012),避免产能不足或产能供过于求的风险和产能建置追高杀低的决策陷阱,以提升资本报酬及整体获利(简祯富,2014a)。

问题与讨论

1. 试举出一实际案例说明数据挖掘和制造智能方法在半导体产业以外其他产业的制造管理上的应用。

2. 请说明美国制造业复兴中,制造智能和大数据分析能力在创新制造扮演的角色。

3. 请探讨德国提出的"工业 4.0"(Industry 4.0)中,制造智能和大数据分析能力的重要性。

4. 晶圆允收测试(WAT)为半导体制程完成后对晶圆所做的电性测试,一般来说,不同产品之间的电性表现皆会有所差异。请利用附件数据 WAT-1.csv, WAT-2.csv, WAT-3.csv(请于本页二维码中下载)分别为三组不同时刻所搜集的 WAT 数据,请由数据面将晶圆数据进行分类,分析各数据集分别包含多少种类的产品?

5. 半导体晶圆加工时,常有某些加工的参数水平或参数水平的组合会对晶圆良率造成影响,假设所搜集的数据如附件数据 Process-1.csv(请于本页二维码中下载)所示,其中,y 字段表示晶圆良率、$x1$~$x10$ 表示各参数水平,请由数据面分析和那些加工的参数水平(或水平组合)会对晶圆造成影响,使良率下降?

6. 半导体晶圆加工时,常有某些加工的参数水平或参数水平的组合会对晶圆良率造成影响,且这些影响具加成效果;例如,某一变量的一水平会造成 2%的良率下降,另一变量的一水平会造成 3%的良率下降,若一晶圆同时使用此二水平加工,则其良率会下降 5。附件数据 Process-2.csv(请于本页二维码中下载)含有上述之加成问题水平,其中,y 字段表示晶圆良率、$x1$~$x10$ 表示各参数水平,请由数据面分析和那些加工的参数水平(或水平组合)会使晶圆使良率下降? 下降多少?



第 13 章

数字决策及商业分析与优化

13.1 决策信息系统

13.1.1 决策信息系统

因应各种问题类型，对于数据挖掘与大数据分析技术，以及数字决策的支持能力也有不同的要求。辅助决策过程的信息整合、决策分析和优化能力的“**决策信息系统**”（**decision information system**）应具备的整合能力，必须符合正确性、稳定性、弹性和容易使用等特性，以加快数据处理的能力和速度。决策信息系统包含信息系统的硬件、软件以及系统的架构、输出入接口和内建的决策模式等数字决策平台，且具有以下优势：

- **快速正确的计算能力**：决策信息系统可以超越人类信息处理的限制，协助决策者快速进行大量的结构化（structured）和非结构化（unstructured）的数据处理与数值计算，迅速正确地产出需要参考的关键指标，提高决策者的反应速度和生产力。
- **更好的信息储存功能**：决策信息系统可以克服人脑储存和搜寻信息的限制，结合数据仓储、数据超市、云计算（cloud computing）以及知识工程与知识管理等技术，提取各种专家的知识和判断并储存于决策信息系统中，并提供快速存取、知识管理与应用的便利性。
- **知识资源共享功能**：借由系统化组织内相似的决策过程和优化方法，决策信息系统不仅能累积及分享每个专家的知识和经验，亦可降低对个别专家的依赖，使专家得以专注于解决更重要的问题。
- **降低决策沟通成本**：决策信息系统具有汇整信息、匿名效果、允许多人同时表达意见等功能，因而得以降低决策参与者在不同时间地点下的沟通成本。
- **清晰明了的结果呈现**：决策信息系统可以通过良好的人机接口设计，如互动设计和图形化接口，协助专家或决策参与者更易于输入、解释与评估相关信息。
- **知识管理外显化**：决策信息系统可以将累积决策分析的经验和对结果的诠释等各种案例，以系统地管理决策相关的决策元素及信息，将内隐的知识外显于知识库平台，作为知识管理的基础。

决策信息系统依照解决的决策问题的特性，基本上分为专家系统、主管信息系统以及决策支持系统等三大类（简祯富，2014b）。

专家系统（**expert system，ES**）内建优化决策模式和算法，作为搜索最佳解（或近似最佳解）的机制，提高结构化决策问题的效率和效果。结构化决策问题通常具有确定的决策目标

与评估标准,因此理性决策者所做的决策应该相同,特别是例行性及重复性的复杂求解问题,例如生产排程等问题,可借由专家系统建立系统化的解决程序。

主管信息系统(executive information system, EIS)提供实时、翔实且多面向的信息,协助决策者处理非结构化的策略决策,决策标准也会因人而异、因时制定。许多大数据数据具有非结构化的特性,问题的结构通常模糊不清或错综复杂,因此需要借助复杂的数据处理方法,高度依赖决策者和专家的主观判断。

决策支持系统(decision support system, DSS)是辅助管理阶层和决策者制定决策的分析工具,经由建立决策规则与模型,将情报搜集、方案产生或方案选择的过程模式化,以提高决策的效能,让客观的信息得以正确地呈现,并协助决策者做出符合个人主观偏好的理性决策,提升计算机在组织中的应用层次,从传统的电子数据处理到协助中高阶层管理者制定日常的决策。决策支持系统将重点置于组织的半结构(semi-structured)或是非结构决策问题辅助工作,融合客观信息与决策者主观的判断,注重决策效益及弹性。决策支持系统强调的是"支持",并无法取代人们做决策,或自动处理决策问题。另外,针对多个决策者的群体决策,则可借助**"群体决策支持系统"(group decision support system, GDSS)**。

决策信息系统发展到现在已可整合各种新发展的技术,例如数据挖掘、云计算、机器学习、人工神经网络及大数据分析方法等工具来强化提供信息的质量和价值,以及用户接口的便利性,其使用者包含了企业内部不同阶层以及不同工作性质的管理者,可应用的范围与领域也越来越为广泛,例如医疗决策支持系统或生产决策支持系统等。

尽管学者对于决策信息系统的观点各不相同,但是大多数都会强调以下特点:

- 为针对半结构或是非结构问题开发而成的决策信息系统。
- 用来辅助而非取代决策者的系统。
- 利用系统化的分析结构来协助与克服决策者的认知限制。
- 利用分析模式与推论功能、数据分析等来系统化评估方案。
- 配合决策思维,引导决策者按部就班地执行每一项步骤。
- 易于使用,且多为交互式接口,可以协助提取决策者主观的判断。

不同类型的决策信息系统,支持的决策问题需求不尽相同:专家系统着重于最佳解模式的建立;主管信息系统着重于信息的完整与实时提供;决策支持系统则强调客观信息与主观判断的结合。决策信息系统的开发与相关的信息技术议题,有兴趣的读者可以进一步参阅(Turban & Aronson,1998)、(陈鸿基,严纪中,2004)。

13.1.2　决策信息系统的架构

决策信息系统包含以下几个子系统,其系统架构如图 13.1 所示。

- **数据管理子系统(data management subsystem)**:即数据库,由数据库管理系统(database management system)软件所管理,其内容涵盖为了解决特定决策问题所需的相关数据。
- **模式管理子系统(model management subsystem)**:为一套软件程序包,提供相关的计量工具以及适当的管理软件。同时亦包括了造模语言,可用来构建特定的模式。
- **知识管理子系统(knowledge management subsystem)**:此系统可以独立运作或是支持其他子系统,并且提供相关程序来增加决策者的智能。

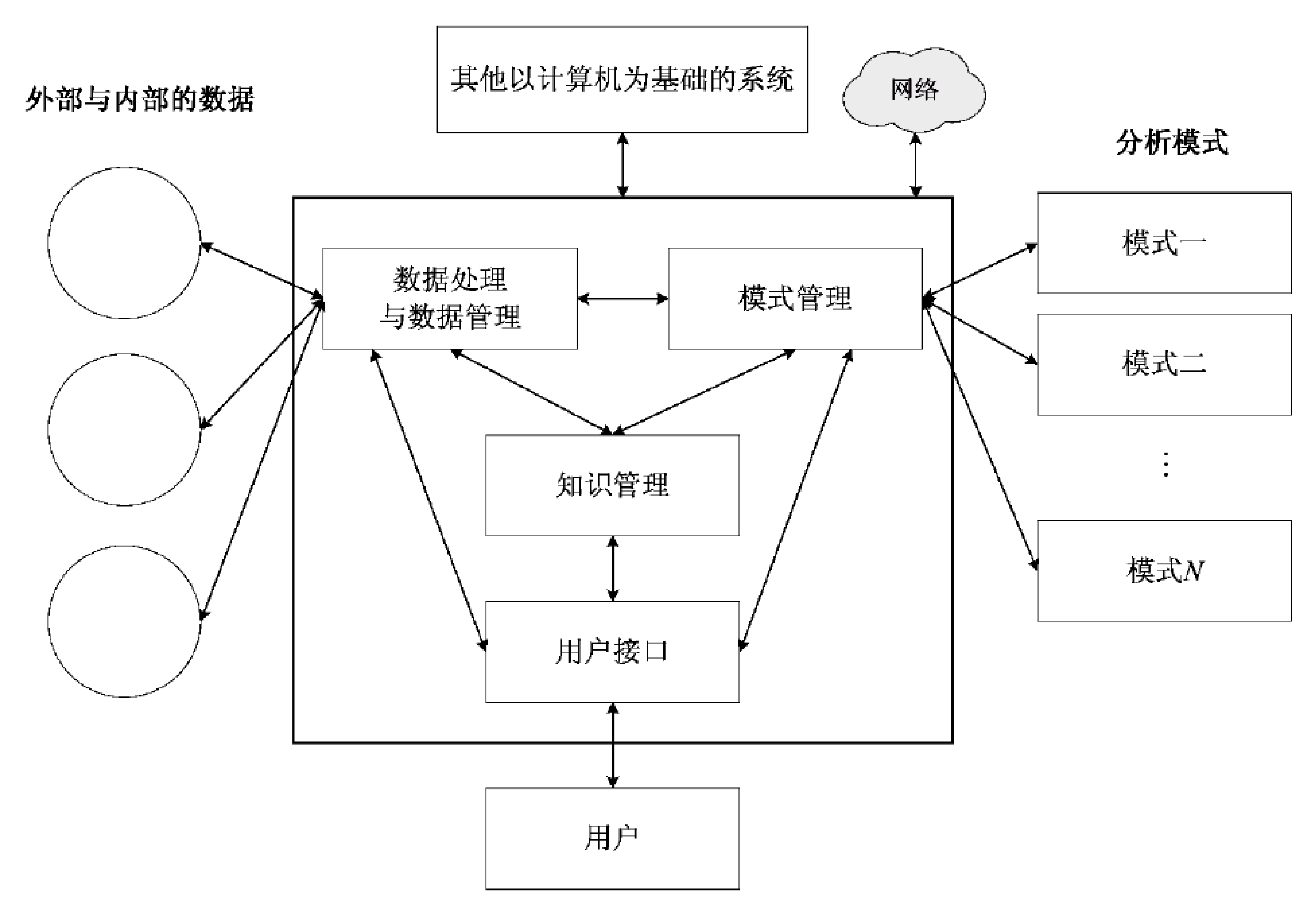


图 13.1　决策信息系统的典型架构

- **用户接口子系统**(**user interface subsystem**)：用户必须经由此子系统与决策支持系统作沟通。
- **使用者**(**user**)：由于人的使用才使系统具有一定意义，因此用户也是系统的一部分。例如，决策支持系统需要借着人机互动的过程来加入决策者的主观判断，用来完成决策的非结构性部分，并结合客观的数据与分析，以形成一个完整的半结构化问题的决策分析。

决策信息系统包含各种能够支持决策过程的系统，利用信息科技、大数据分析方法及商业优化模式，将决策问题中可结构化部分加以分析，并建立架构或模式；同时在推导的过程中引领决策者加入非结构化的判断，并提供多个方案或建议供决策者选择。决策信息系统是具有推演分析、比较可行方案、寻找最佳的建议等功能的人机交互式(interactive)系统，以帮助决策者提升决策绩效。

13.1.3　应用实例——电性测试机台维修的决策支持系统

1. 案例说明

集成电路(IC)组件的终端测试(final test)常会因为机台故障而导致测试结果偏离正常值而必须进行机台维修与保养以及重测的问题。当一批 IC 组件投入分类机(handler)中进行测试时，如果有些测试管(site)的良品产出率明显低于其他管，或是有些分类机的良率相对较低时，工作人员就必须评估低良率(low yield)问题是源于测试机台本身故障而造成测量结果的不准确，或是产品质量的异常。当确定是测试机台的问题时，工作人员就必须决定是否将异常的测试管关闭，或停止整个测试工作以进行维修与保养。本案例架构了系统化的决策分析流程，以不断厘清问题的本质，了解决策元素之间的关系，利用图形化以及方程

式架构出整个问题的核心,并厘清目标与建立目标方程式,以分析在每个批量中断时点所做出来的不同决策的影响(Chien & Wu,2003)。

2. 架构决策模式

测试机台维修决策可以架构成半结构化的决策分析问题,决策者为现场作业人员。在测试因故暂停,例如两道测试(N0 与 N1)的间隔、人员交班或机台卡料时,作业人员会将即时机台的测试状况储存,称为批量中断(lot end)。如果仅考虑正常测试过程中的批量中断,则可分为三种情况:①同批 N0 中的批量中断;②同批 N0 与 N1 之间的批量中断;③不同批之间的批量中断。三者在测试流程中的位置如图 13.2 所示。

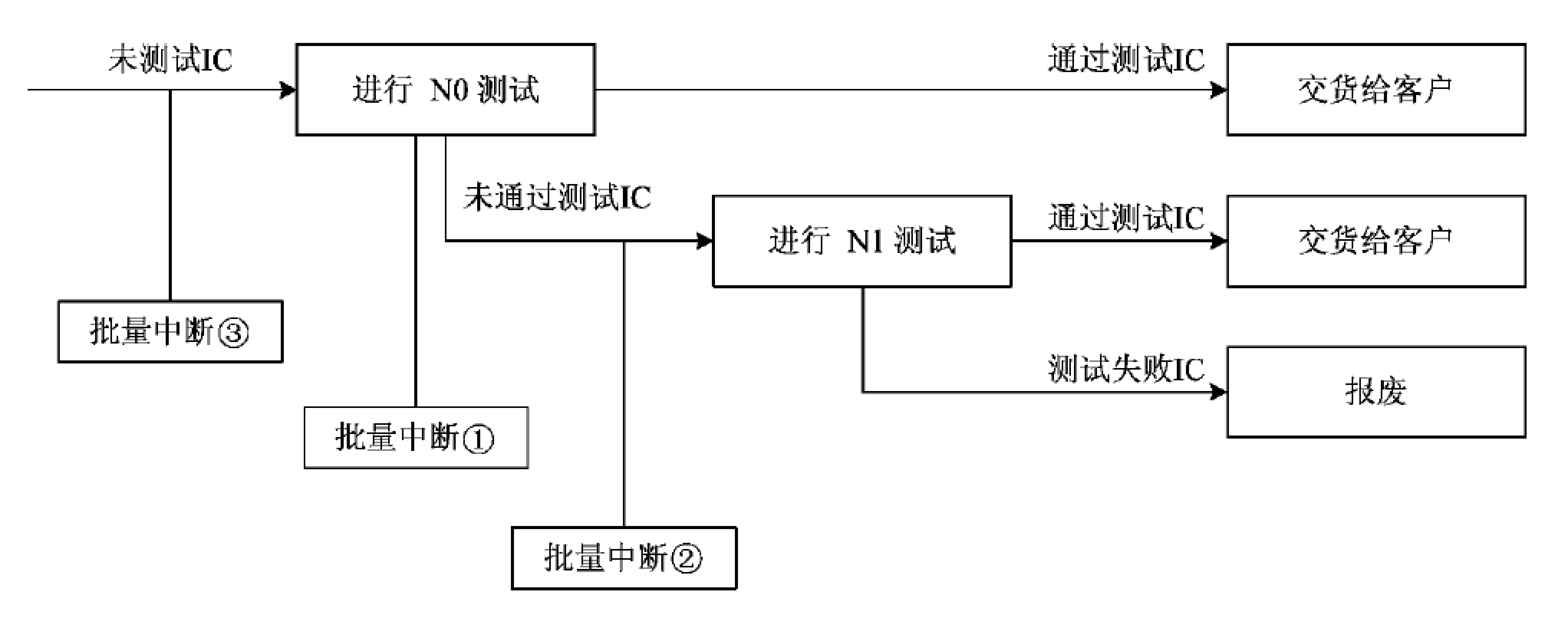


图 13.2　电性测试流程图

批量中断后会根据计算机读取测试状况的结果,作业人员必须依照目前分类机中各管良率差异的变化,进行关管、关机维修或继续测试的决策。其中,关管为决定哪个测试管不进行测试,并由良率低到高依序关闭。因考虑良率太低的测试管投料可能会因为测试出良品的颗数太少而被要求重新测量,导致额外的测试机台使用时间。关机维修可以使良率太低的测试管恢复到正常测量功能,只是整个主系统都必须停止测试而降低机台的有效利用率。同时,机台维修所需花费的时间不一,必须进行预修之后才能进一步评估,并且根据估计维修时间的长短来决定是否继续维修,或者只是关闭异常的测试管,整体流程如图 13.3 所示。

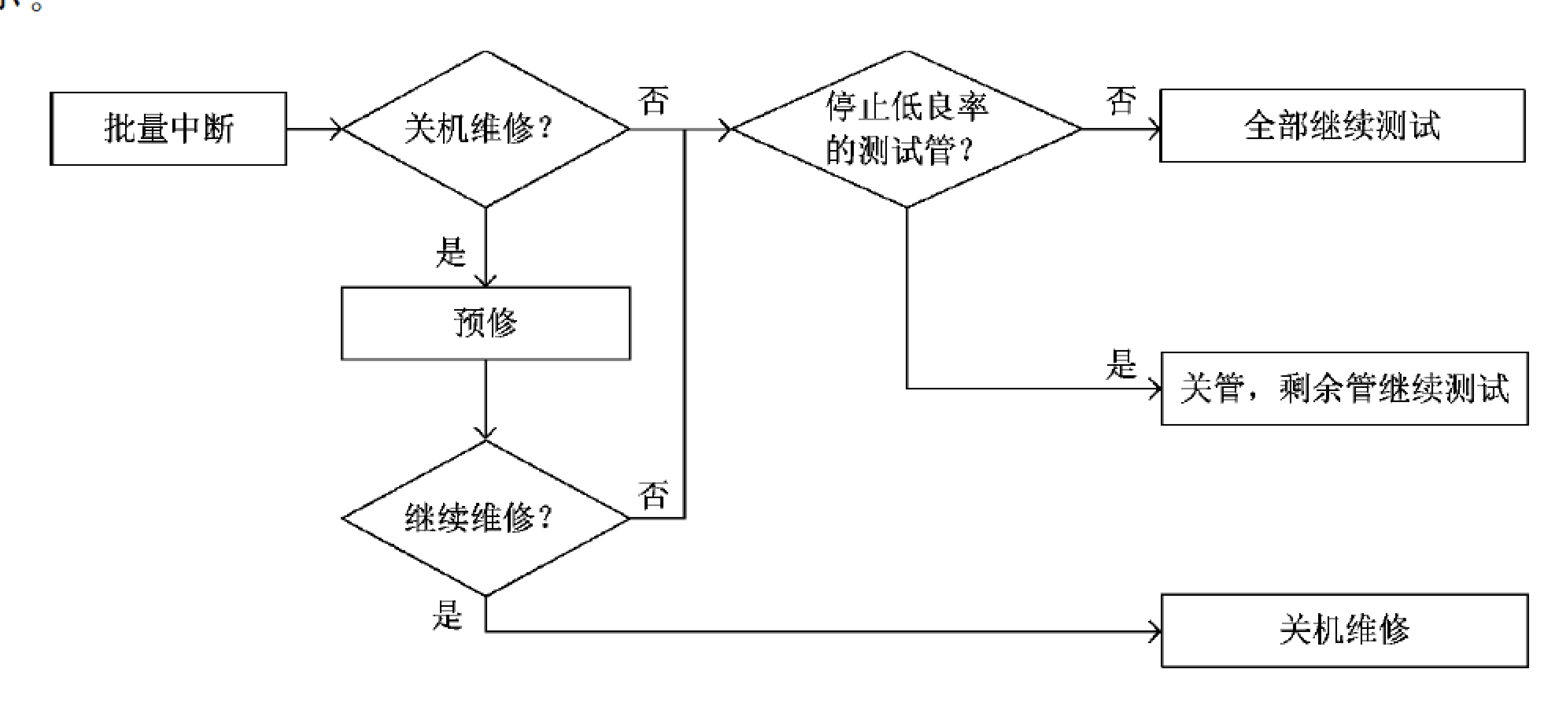
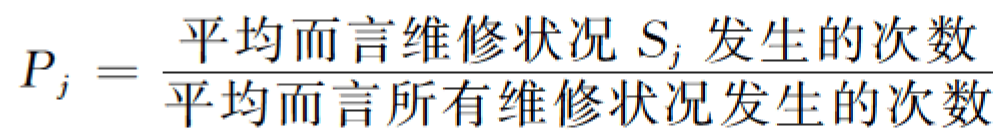


图 13.3　批量中断后须实行的决策流程图

以决策树来架构此决策问题如图 13.4 所示。首先第一层的维修决策(D_1)乃是决定关机维修(B)或是不关机而开 n 个测试管继续量测(A_1, A_2, …, A_n),其续测的获利值以 $F(A_n)$表示。决定关机维修后则会有预修的动作,需维修的时间为 $\bar{s}$,预修后则可进一步评估还需要多少时间才能够维修完毕。但是此剩余的维修时间是不确定的,可视为是一个机会点,每个机会点后的分支为各种可能的维修情况(S_j)的实际剩余时间 s_j。其发生的概率可由历史数据推算如下:



$$P_j = \frac{\text{平均而言维修状况 } S_j \text{ 发生的次数}}{\text{平均而言所有维修状况发生的次数}}$$

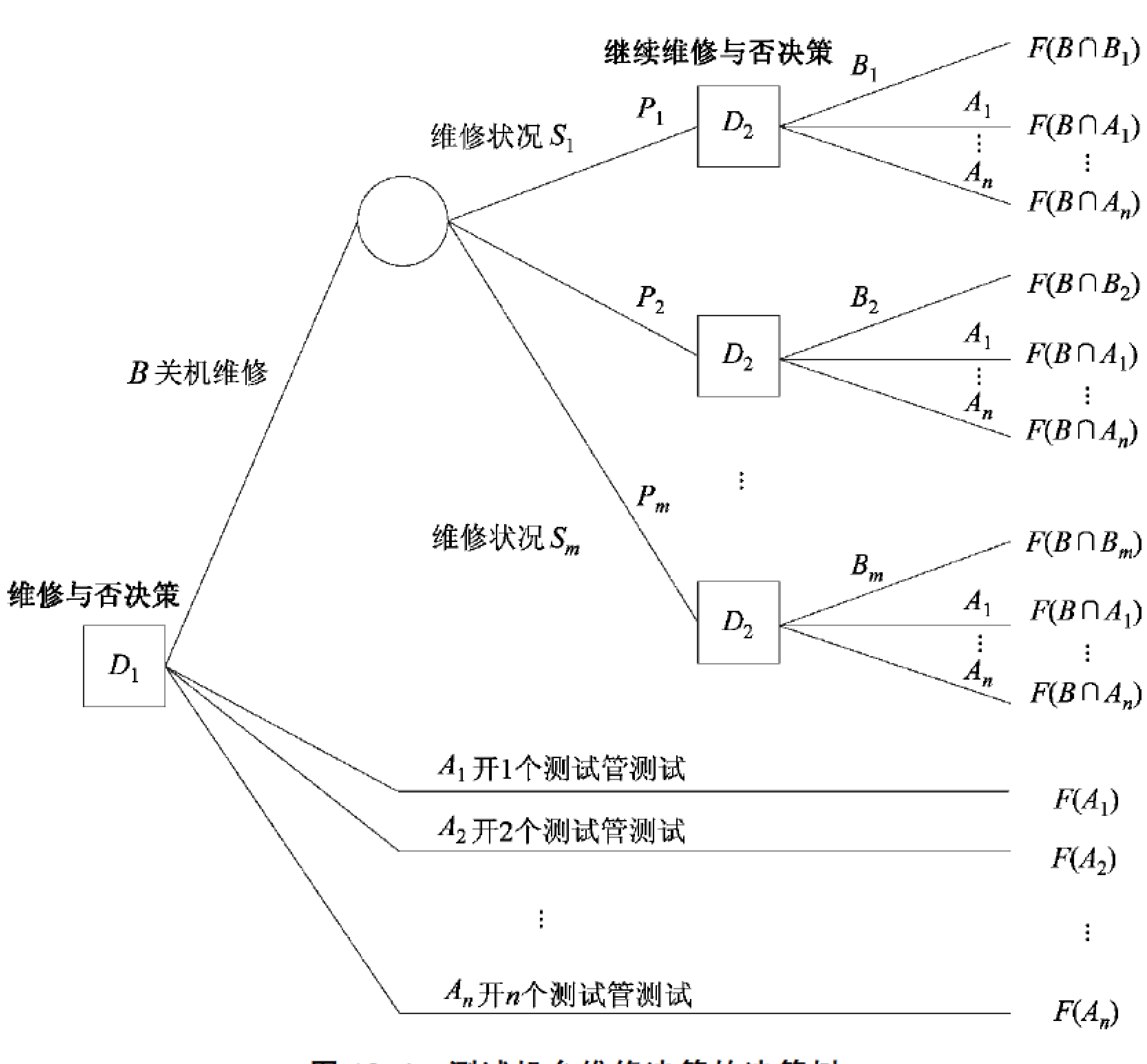


图 13.4 测试机台维修决策的决策树

假设有 m 种的可能维修状况,$j\in\{1,2,\cdots,m\}$。每一种维修状况之后接续第二层维修决策(D_2),此层决策为根据已知剩余的维修时间来判断是否继续维修或续测。若决定继续维修,则获利值为 $F(B\cap B_j)$,$j\in\{1,2,\cdots,m\}$,维修时间为 $\bar{s}+s_j$,若决定续测,则获利值为 $F(B\cap A_i)$,$i\in\{1,2,\cdots,n\}$,编码 i 是依照试管良率由高至低排序,而关管顺序是由良率的低至高。根据不同维修状况下进行的不同决策,决策者可以求得机会点的期望获利值 $\sum_{j=1}^{m} P_j \times F(D_2 \mid S_j)$,并将此期望值与不关机决策的获利值相比较,以选择维修决策的最佳决策方案。

本案例的决策目标包含四个目标的量化衡量,P 为通过测试的良品可获得的利润,良品通过颗数越多则获利越高;T 为测试时间对应的测试成本,整批测试时间越少则测试成本越低;D 为交期延迟的过期损失;Y 为未达正常良率水平的惩罚金额。以这四个目标建立此决策问题的目标方程式,决策准则为最大化利润,即总利润扣除测试成本、过期的损失以及低

良率的惩罚金，如式(13.1)所示：

$$F(n_0,n_1 \mid c,v,w,b,d,t_s,s,R,Q,Q_p,Q_f,r) = P - T - D - Y \tag{13.1}$$

其中，由于案例公司无法将每批待测组件的到期日分摊到每个测试站，而且在良率水平方面并没有一定的准则与依据，因此式(13.1)的过期损失 D 以及未达良率水平的惩罚 Y 被予以省略。而总利润 P 等于单颗良品组件的利润 c 乘以总测试颗数，测试成本 T 等于单位时间的测试机台成本 v 乘以总测试时间，定义如下：

$$P = c\left(\sum_{i=1}^{n_0} u_i Q_0/n_0 + r\sum_{i=1}^{n_1} u_i Q_1/n_1\right)$$

$$T = v \cdot t_s \cdot (Q_0/n_0 + Q_1/n_1)$$

n_0 与 n_1 分别为测试 N0 与 N1 的开管数，Q_0 表示此决策点后所需要的测试组件总颗数，Q_1 表示决定关管数 N0 后 N1 仍然需要测试的颗数，因此 Q_0/n_0 与 Q_1/n_1 即代表 N0 与 N1 各测试管所装载的待测组件颗数；而 u_i 或 ru_i 则为各测试管的良率，因此通过测试的组件总数可表示为 $\sum_{i=1}^{n_0} u_i Q_0/n_0 + r\sum_{i=1}^{n_1} u_i Q_1/n_1$，也就是各测试管的待测组件的颗数乘以良率；而 t_s 为单颗组件的测试时间，因此总测试时间可以表示为 $t_s \cdot (Q_0/n_0 + Q_1/n_1)$，也就是待测组件颗数乘以测试时间。改写后的目标式如式(13.2)所示：

$$\text{Max } c\left[\sum_{i=1}^{n_0} u_i Q_0/n_0 + r\sum_{i=1}^{n_1} u_i Q_1/n_1 - vt_s(Q_0/n_0 + Q_1/n_1)\right] \tag{13.2}$$

另一方面，可以比较关机以及不关机等方案的目标值，然后求得可以权衡的最大可接受的维修时间(acceptable repair time)。当维修人员被告知维修时间之后，就可以判断自身的技术与经验是否能在此时间内完成，若有把握完成维修则关机，否则继续测试。在实务中，测试机台在测试不同的产品项时会有架机(setup)的动作，此一动作会同时将各测试管中不良的状况排除，而决策点至下次架机间所需测试的颗数(Q_s)，可以经由每天的排程估算出。

接着令 $Q_0 = Q_s$ 代入式(13.2)，并且经由比较关机以及不关机的获利来求得可允许的维修时间 s。式(13.3)表示关机以及不关机的获利比较式，等号上方为开机测试中的最佳决策目标获利；等号下方则为关机维修的目标获利，其中，$\bar{u}_i$ 代表各测试管在维修之后的良率，N 表示总测试管数。

$$\begin{aligned}&\text{Max } c\left(\sum_{i=1}^{n_0} u_i Q_0/n_0 + r\sum_{i=1}^{n_1} u_i Q_1/n_1\right) - vt_s(Q_0/n_0 + Q_1/n_1)\\ &= c\left(\sum_{i=1}^{N} \bar{u}_i Q_0/N + r\sum_{i=1}^{N} \bar{u}_i Q_1/N\right) - v\left[t_s(Q_0/N + Q_1/N) + s\right]\end{aligned} \tag{13.3}$$

3. 架构决策支持系统

为了协助联机操作员进行复杂的关管或关机维修决策，本研究发展内建上述决策模式的决策支持系统功能，并内建在测试机台中。当作业员开始执行批量中断时，决策支持系统即可根据最新的测试数据来测试目标式并且建议是否应该继续测试、关闭部分测试管或停机维修等。当系统提出关闭部分测试管的建议时，会依据测试良率的高低来建议关管顺序。另一方面，若系统建议停机维修，则会估计出最大可接受的维修时间，让设备工程师判断是否有把握在合理的时间内完成维修才停机或继续测试而仅关掉部分测试管。其决策支持系

统的架构根据图 13.1 所制定，如图 13.5 所示：

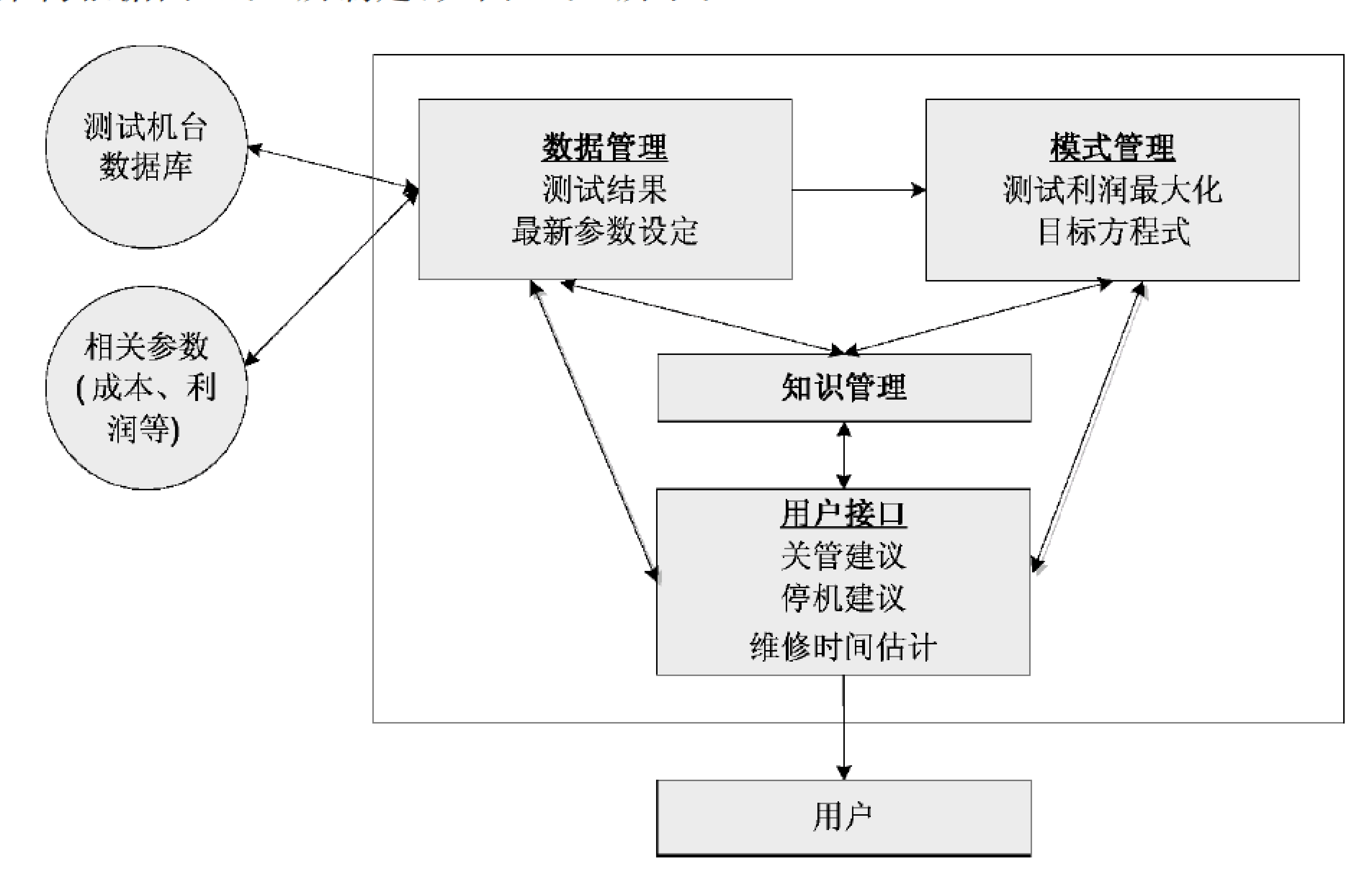


图 13.5 测试机台关管、停机维修的决策支持系统架构图

4. 结果与讨论

本案例针对有 32 个测试管的测试机台，经由模式计算，建议将测试管 1 与测试管 7 关闭，预估可测试通过 319 颗良品。以历史数据来说，作业员关闭的测试管共有 14 个，实际通过良品的颗数为 323 颗，比模式决策结果所预期的良品多 4 颗，然而因为关闭的测试管过多，故实际完成测试的时间约为决策模式的两倍。

分析实际作业人员所做的决策发现：①作业人员并未将所有的待测组件平均分配在所有已开的测试管中，例子中的测试机台为装载两台分类机的主系统，作业人员将整批待测组件不平均地分配到两个分类机中，以至于整批完成的时间受到测试较多颗的分类机影响而膨胀；②作业人员将良率较高的测试管关掉，然后让良率较低的测试管续测因而导致整批的良品颗数偏低。同时，关管数过多致使整批完成时间偏高；③作业人员的决策明显没有考虑到良品颗数、测试时间与这两个属性之间的关联性，因此所做的决策以这两个属性来衡量均为不佳的决策。

本案例可以提供联机操作员实时决策的有效方法，提供决策支持系统的解决方案与决策的规则，并且进一步考虑权衡测试产出、测试时间、准时交货以及测试质量等目标。中国台湾由于缺乏先进半导体机台设备制造商，使得许多半导体厂改进生产流程和提升良率的许多参数设定和调整，往往在国外设备厂商进行保养时就可以发觉，而有可能外流给设备商的其他客户，无形之中助长了潜在对手跟进的竞争力(简祯富，2014a)。通过结合信息系统和大数据分析所打造的制造智能平台，可以将公司所累积的分析知识和重要诀窍储存在内部的云端知识系统，避免流失的风险，使产品良率、产能利用率、生产效率、机台妥善率等得到惊人提升。

运用大数据分析提升高科技产业的制造智能是台湾半导体产业未来能否持续领先的关键。过去，受限于硬件技术，使计算机的运算能力不足以符合“实时决策”(real time decision)的实际应用需求，因此实时决策系统目前仍然以学术探讨为主，而尚未真正在产业

界落实。但在硬件技术的长足进展下,现今计算机的运算能力已显著提升,大幅缩短决策系统的处理速度,"实时决策系统"在产业界的应用将会日趋普遍。尤其对制造结果精确度有高度要求的半导体产业技术蓝图已将机台设备的实时决策能力列为发展重点。台湾高科技产业在自动化制造和检测过程中,累积了庞大数据,由于数据的变动性,这些数据若未能实时有效分析,只是花钱买设备系统储存而未善加利用,不仅不能成为资产反而是企业的负债。所以,若能导入大数据分析技术,从中挖掘潜在有用的信息,将是料敌机先的制胜关键。

13.2　商业分析与优化

13.2.1　商业分析与优化

金融海啸使得全球的经济环境产生剧烈的变化,新的商业模式和全球化正逐渐影响所有企业。企业运营过程中其实隐藏着大量有价值的信息。例如,管理者想知道某个关键供货商的存货水平,或更进一步了解顾客的购买行为、更清楚合作伙伴的运营与财务状况,以及预测未来可能会对于企业运营环境造成影响的事件。善用**商业分析与优化**(**business analysis and optimization, BAO**)的企业可以从多种角度来获得即将要面对的问题,深入了解顾客需求,并且更有效地预测供应链的限制以及竞争对手的应对方式,快速制定决策,领先竞争者并面对后续的挑战。

过去 30 年来,计算优化(computational optimization)技术进步三千多倍,计算机运算能力达万倍,整体分析能力增强三千万倍以上(Bixby, 2002),且非结构性数据的处理需求也长足进步,让商业分析与优化应用条件日益成熟。简祯富在接受《IBM 蓝色观点》(2011)访问时,就指出:没有经过分析与优化的决策,可能导致方向错误,达不到预期的结果。台湾企业非常弹性,使瞎忙的成本被忽略,决策效益也未受到适当的检验。因此,企业各部门的管理阶层必须转变观念,从过去专注于成本管控转而成为提供信息、分析与决策的价值整合者,以协助公司迈向智能型企业。

商业分析与优化指的是善用决策分析复杂环境与数据,并且为企业找到最佳的方案来优化资源运用以提升企业价值的能力,亦是未来企业经营决胜的核心能力。其可使管理阶层更系统化地分析复杂数据以提供决策者所需的信息,接着建议可行的最佳方案供决策者参考,使决策者能投注更多心力思考企业的策略规划以改善决策质量,将资源用在更有效益的地方。

商业分析与优化的主要目的即是提升企业的分析观察能力,其具体功能与实施步骤可分为以下三个阶段:①规划未来信息:由企业提供符合目标的策略性项目,有效地应对未来各种变化;②管理信息:确保信息的准确度、攸关性以及安全性,为企业带来数据完整性以便于管理、运用、分享以及再利用;③优化商业分析:从广泛且互相关联的信息中有效地进行观察,并通过分析预测其商业价值,辅助企业策略出最有效的决策方案。

商业分析与优化是大数据时代中,未来企业决策者进行决策制定的重要辅助工具之一。IBM 针对全球企业领袖的调查发现,有 1/3 的企业领袖经常被迫在信息不足的状况下制定重要决策;而有 1/2 的企业领袖经常无法获得充足信息;IBM 的调查报告(LaValle *et al.*, 2010)也指出,企业面临的最大挑战中,"如何创新以达到差异化"(61%),更胜于"提高营收"

(50%)及"降低成本、增进效率"(46%)。因此,数据挖掘与大数据分析具有发掘潜在未知可能的洞悉能力,就显得十分重要。许多高绩效表现的企业,都一致认为巨量数据分析和决策是未来达到竞争优势差异化的关键能力。

然而,台湾大多数企业的领导者或高阶主管,仍然欠缺对决策分析、数据挖掘和大数据分析等技术和工具对企业经营的影响的了解,这也是许多企业的公司治理无法从人治走向科学管理与数字决策的关键。换言之,采用商业分析与优化系统的障碍,主要来企业内部对于使用分析系统来改善企业营运缺乏足够的认识,因此,决定企业采用分析系统与否,主要和企业管理、组织架构、决策过程与企业文化有关,而不是数据或技术问题。

尽管大多数企业已采用统计分析和全面质量管理等方法来协助分析相关的运营数据,但在大数据时代数据快速累积、变动、非结构化的特性,使传统的统计分析方法不再符合企业所需。现今企业需要更有效的方法与分析技术来进行商业分析与优化。

13.2.2 商业分析与优化的基本要素

商业分析与优化可以从冗余的大量数据中解决信息过量的问题,帮助企业更完善地制定公司决策,应用于各行各业之中。例如,客运、物流等交通运输业者可以借由实时预测主要干道的交通变化模式,估计出最省时省油的路线,同时达到节省成本与客户服务的目标;移动通信业者可以分析用户实际的社交网络使用模式,归类出不同的客户群,针对个别客户群设计不同形态的资费方案;医院可以通过分析由先进传感器和传统监控装置所实时监测与持续搜集的详细生理数据(例如,心跳与呼吸频率),提早推测出可能出现的感染或潜在疾病。这些应用都是通过大数据分析所导引出来的新行动方案与模式。

商业决策的优化需要分析各式各样的目标、限制与内外部资源,并将优化的决策落实于企业运营中(LaValle, 2009)。举例来说,移动通信业者可能从数据分析中发现某种样型或规则,进而决定推出崭新的客户服务策略与运营模式。例如,在新产品或新技术推出时,优先推广给偏好新科技的客户群,或推出不同的组合方案服务给可能带来较高获利的客户,但另一方面,也持续维护低贡献但高忠诚度的客户群,衡量企业运营资源以优化其运营方式。针对不同类型的客户提供定制化的服务项目,维系客户的忠诚度与使用意愿。

另一方面,通过整合大数据分析、数据挖掘与统计分析等方法,以预测可能发生的事件并预先做好准备,主动评估成果并权衡利益得失,在市场情况变化之前主动预测与规划资源和优化结果,将企业调整成能够达成新目标的最佳状态。包括提供商业分析与优化策略,以解决方案的分析能力加强企业对公司各项业务的掌握程度,包括对消费者、市场、竞争对手的观察等,协助管理者降低风险、减少成本,以较快的速度来达成商业目标,例如,Chien 等(Chien *et al*., 2010)发展半导体需求预测技术,以协助产能规划与建置决策。

企业不断地追求更精简的供应链、更好的顾客服务以及更快的时间内获得更高的利润,然而,现今企业永续经营已无法仅依赖效能的提升。大多数的企业都会将资源投入在提高自动化以及效率方面,使用相似系统和流程去产生类似的产出结果。然而,真正的观察力是要使组织易于了解顾客的需求,进一步从事风险管理与资产管理,并且要想尽方法不轻易地被竞争对手复制。信息是许多成功企业内部一项非常重要的资产,采用强而有力的管理信息平台,能比其他企业都更有能力掌握信息,建立决策型组织(decision organization),制定正确的决策与发挥信息最大的价值,为企业带来良好的经营业绩(Blenko *et al*., 2010;

LaVallе *et al.*，2010)。

事实上，许多企业已经开始以个案分析作为“问题点”的突破来产生最佳解决方案的路径，提升企业自身的竞争力与商业利益。商业分析与优化所带来的效益显示，企业必须使信息分析和管理成为公司内部不可或缺的一环，并将数据视为重要资产，重新定义企业优化的需求，并设计相对应的信息策略与信息平台，以达到优化的目标。

13.2.3 商业分析与优化的应用

商业分析与优化的目标是要让企业依照最佳的行动方案的基础上能实时获得相关的观察与可靠的信息，进一步推动决策，增加组织的能见度，并且运用这些能力以找到更好的顾客(IBM Corporation，2011a，2011b)。

1. 提高对顾客的理解程度

更深入地了解顾客的消费行为与喜好等，借以规划不同的营销方式或更具有吸引力的产品，掌握购买趋势与交叉销售的机会，增加顾客或预防顾客流失，以维持企业获利。例如，提供在线影音服务的 Netflix 借由大数据分析过去客户收看的习惯与记录，找到喜欢观看《纸牌屋》的观众，有一定的比例也喜欢导演大卫・芬奇(David Fincher)与演员凯文・史派西(Kevin Spacey)的影集，Netflix 也因此决定主动出击投资《纸牌屋》的重制与拍摄。不仅如此，顾客行为分析可以进一步掌握促销活动的时机与市场趋势，并通过将每一天的销售转换成企业的预测和分析模型，借以改善生产与销售的能力。企业本身也可借此控制营销成本，减少浪费性的支出与不相关的优惠。例如，随着时间增加与记录的累积，Netflix 也借由顾客的收视分析，对于不同影视类型的评价，提高影视节目推荐成功的概率，使得 Netflix 得以减少营销上的开销。

2. 实时导引出优化决策

大多数公司业绩报告、商业预测都是依赖历史数据和固定的流程而建立。然而，在信息随时都在变动更新的大数据时代，这种回溯性的数据分析已不足以达到最佳的商业决策。企业必须以“实时分析”来创造差异化与优势，导引出优化决策。例如，抢先竞争对手了解市场状况和顾客的需求(Chien *et al.*，2010)，以确保能够从供应链获得较好的价格与稳定的供应；根据最新的市场滚动需求预测信息，实时调整公司产能规划决策，降低产能过剩或供给不足造成的损失(Chien *et al.*，2012；Chien & Zheng，2012)；使用模式的分析和预测的分析可以实时监测企业内部可疑的活动，并且及时在亏损发生之前采取适宜的行动；对于着重于时间因素的决策，如医生能够利用系统进行重症病人的治疗，或是使用远程监控数据来查看最新的信息并进行购买决定，商业分析与优化都能使企业的信息在竞争活动上更具有灵活性。

3. 跨组织整合决策

跨组织整合决策能将关键信息分享给所有的利害关系人，使企业各层级的员工能实时获得相关信息，也可以使企业更准确地依据信息优化相关决策。特别是大型复杂的组织，更需要通过商业分析与优化来整合相关信息。以改善顾客服务为例，建立一致的服务信息可以给予顾客更佳的服务体验，同时降低成本，减少在互动时所需要解决的问题。例如，简祯富提出 PDCCCR 制造策略架构(Chien *et al.*，2010)，以整合跨组织的定价(pricing)、需求

(demand)、产能(capacity)、资本支出(capital expenditure)、成本(cost)和收益(return)的相关决策。

4. 确保企业运营保持最佳状态

提供全面信息管理与商业分析能力,可以在复杂的环境中,挖掘出原始数据的背后隐藏的信息,以在变化多端的环境中保持领先地位,其所涵盖的层面包含:①策略校准:通过关键性指标,企业可以随时检查、调整以及优化其经营策略;②风险管理:借由广泛性的企业分析报告,可以全面且实时地让企业决策者随时知道公司的状态,并且采取适当的策略;③需求管理:通过准确的预测更能平衡供给与需求,有助于降低存货成本以及优化资源分配。另一方面可以使企业比其他竞争对手更迅速地朝向新兴市场迈进。

13.3 数字决策

大型且复杂的决策问题可能会包括许多的决策元素,也使得问题的组合复杂度呈现几何级数增加,其中包含多个不确定事件、多个属性、多个方案甚至是多个决策者等。由于人类大脑的认知能力和对信息的处理能力有其上限,因此巨量的数据搜集与复杂分析以及各种决策模式的建立,须借助信息科技以提高决策速度与决策质量。此外,如果类似的决策问题会重复在不同的时空中发生,借由信息科技将决策过程与方法标准化、系统化,更可不受时间的限制而持续使用。

利用信息科技以协助企业进行数字决策的决策支持系统,以及借由商业分析与优化为企业分析找到最佳方案来优化资源运用提升价值,着重于如何以信息科技来构建系统化的决策分析过程,然后纳入人类专家知识于决策过程中,以克服人类心智的限制,让决策者进行决策的同时也能够兼具信息科技与人类专家知识的优势。

随着企业运营环境的变动日益快速,市场竞争日益激烈,现代决策者必须准确地进行决策,才能因应不断更新的顾客需求以维持竞争优势。由于信息科技的进步,使得决策者可以借由各种信息系统来辅助决策,通过创新的数据挖掘、决策分析与优化方法,以便更精确地掌握关键信息,快速地进行复杂的分析与执行系统化的评估,帮助我们建立主动规划的策略和智能型的数字决策,而非反应式行动,来解决复杂的问题并改善运营绩效,获得更具一致性和更高质量的决策。

以高科技制造业而言,在单一机台上各个环节现今多已采用信息系统辅助作为连接并获得良好成果。因此,简祯富在接受《电子时报》访问时指出(谢佩原,2011):智能工厂的制造执行系统(manufacturing execution system, MES)已逐渐走向"智能化",在不同的机台间引入"实时决策系统",建立共同的标准沟通接口,进行整合以自动协调分配各机台资源,将是未来主要的发展方向。但设备仪器越"聪明",工厂的人力需求程度也越低,在生产制造领域的竞争优势将逐渐转移成设备的竞争。届时,企业资本越雄厚,越买得起"聪明"设备的厂商,其效率与竞争力就越好,台湾硬件制造厂商也将逐渐失去目前的领导地位。因此,企业必须掌握转型的契机,基于硬件知识发展软件业,才能开创台湾产业的新局。

另一方面,企业组织也必须转变,让制造过程的信息从自动化到决策化,此一转型的过程是趋势也是挑战。以半导体产业的数据挖掘而言,虽然目前已可以深入发掘各种制造信息,但由于半导体制造程序复杂、影响变量众多,往往无法从搜集的庞大数据中,迅速有效地

挖掘或归纳其中有意义的样式或规则，更遑论提供实时决策的依据。因此，数据挖掘要产生效益，人力资源将扮演重要的关键。主管要根据价值来源来规划组织架构，决定决策"所有权"(ownership)的授权，通过组织层级将人与工作任务展开，使每个人的工作任务与权责都非常清楚，各有决策负责人。结合大数据分析工具、商业分析与优化以及数字决策系统，每个人就可从决策过程中培养判断能力，提升决策能力，主管才能将时间分配至更具前瞻性的策略规划工作上(简祯富，2014a)。

关于辅助决策的工具，从通过商业智能、数据采矿等搜集信息，到提取信息以支持商业决策，IT 都可提供协助，但更重要的是，组织必须跟着动。以前所谓企业流程再造，其实都是借由 IT 进行自动化，将原本机械性的动作改为计算机化，而忽略企业营运更复杂的流程，其实是决策流程。因此，组织与决策流程也必须再造，才能协助企业在决策中有效进行资源分配，让各部门各司其职，发挥综效。

云计算是提供可随时、随地、随选地经由网络存取共享的资源服务，包括运算资源、网络资源、储存资源等，这些资源可在不同用户间动态地分配与调整。一般而言，云计算的模式主要有三种：软件即服务(software as a service, SaaS)、平台即服务(platform as a service, PaaS)、基础架构即服务(infrastructure as a service, IaaS)。SaaS 主要是借由网络方式提供软件服务，例如电子邮件、在线游戏等；PaaS 则是提供企业执行软件运算所需的环境；IaaS 则提供底层数据储存与运算的资源。企业通过云计算服务的提供可直接获取与累积大量的客户数据，如何从中结合大数据分析技术以提供客户更多元化的服务，是未来的重要趋势。

数据传送方式随着传感器、无线网络技术的发展，使得机器对机器(machine-to-machine, M2M)之间的连结越来越紧密，也带来各类各样的数据。物联网(internet of things, IoTs)即是通过无线射频识别标签(radio frequency identification, RFID)、无线网络、传感器技术将物品相互串连，形成一个网络，在此网络中得以随时掌握物品的动向与状态，并自动地提供信号与结合智能化的技术提供实时分析。

决策分析、数据挖掘、大数据分析、商业分析与优化、数字决策在企业的导入和发展是个循序渐进的过程。许多企业刚开始都只有使用商业分析来制作报表，然后引进数据挖掘和商业智能来进行分析，进而导入商业分析与优化以进行决策的优化，并利用预测性分析工具和大数据分析技术来逐步构建数字决策的能力。累积"问题点"的突破，扩大为"系统面"的大数据分析架构与数字决策的完整解决方案，协助企业在大数据时代中保持竞争力。

13.4　结论

面对数字化的新经济模式与智能化的信息环境，企业如何善用数据挖掘、大数据分析、商业分析与优化等工具是决定企业能否在未来继续成长与获利的核心能力。信息负荷过重是制定良好决策的一大障碍，但借由目前的技术与分析专业能力，巨量信息将可以带来真实的利益。信息越密集与多样，企业对未来的趋势预测就越准确，进而采取有效行动以掌握预测出来的各种商机。

能够在未来脱颖而出的企业更愿意挑战现况、积极掌握大数据的潜能，并提供直接接触大数据的各个领域的员工所需的信息工具来实时做出有效分析、建议或决策。这些工具能让企业将信息可视化，并且预测在其他情境下会有什么样的结果。过去的观念认为，决策是

高层主管的工作,基层员工只需执行即可。但公司的组织架构应该是要协助组织做出比竞争对手更好、更快的决策。在网络时代,如果所有人都要坐等高阶主管下决策后才开始行动,反应速度肯定来不及。所以,上下层级分明的官僚结构已不符合现今产业的需求。借由善用决策分析方法和数字决策工具,不仅幕僚可更系统化地分析复杂数据以提供决策者所需的信息、建议可行的最佳方案,决策者更可专注心力思考策略规划以改善决策质量,而减少“瞎忙”的情形产生,并把资源用在更有效益的地方。高科技制造业应思考发展决策分析与优化的制造智能系统,以将最佳的绩效落实到各项决策中,从而优化整个企业的运营绩效。

在大数据时代,每人每天都接收到大量的信息,企业决策者的角色和定位必须改变,需要进行决策流程再造。卓越企业应该是“决策型组织”,使“人人都是决策者”,来提升决策反应的速度和质量。高阶主管要能适当地授权,转而成为制订策略、维持价值和整合决策信息的领导者,让第一线接触数据的员工能够借助大数据分析工具实时做适当的判断,通过层级分析定出组织不同阶层、不同功能的每个人的关键绩效指标和决策所有权。厘清权责,借助巨量数据中的实时信息,使很多事情能在第一时间处理,才能因应网络时代的快速竞争,掌握决策契机(简祯富,2014a)。

问题与讨论

1. 请讨论决策支持系统在大型的企业管理信息系统中,例如供应链管理或是先进规划与排程系统中的角色和应用。

2. 请根据本书所介绍的决策支持系统的特性,探讨决策支持系统的关键成功因素。以供应链管理系统为例,讨论如何构建供应链系统建置的评估指针。

3. 请举出几个数字决策、商业分析与优化的应用实例,并讨论比较其特性。

4. 请说明并讨论数据挖掘与大数据分析在物联网的应用方向,或举实际案例。

5. 附件数据 pattern.csv(请于本页二维码中下载)为 10 000 组图像扫描的参数设定与图像质量历史数据,其中图像质量数据为文件名 Map_X.png 的文件,其所呈现的图案即为图像扫描时缺陷的样型。一般来说,各种缺陷样型可能对应各种特定参数的输入所造成。因此缺陷样型的分类可视为故障排除的第一个环节。

(1) 请由数据中,定义各种不同缺陷样型。

(2) 请找出各种缺陷样型所对应的原因(参数输入范围)。

(3) 在数据所见的缺陷样型的范围内,请试着将分析模式模块化,以构建自动化故障检测模式。



参 考 文 献

中文书籍

MAYER-SCHONBERGER V，CUKIER K. 2013. 大数据[M]. 林俊宏，译. 台北：天下文化.

PAPOWS J. 1999. 16定位[M]. 李振昌，译. 台北：大块文化.

TURBAN E，ARONSON J. 1998. 决策支持系统[M]. 李俊民，编译. 台北：华泰文化.

ZORATTI S，GALLAGHER L. 2013. 标靶营销：中！紧贴目标顾客，省下99%的乱枪打鸟[M]. 杨如玉，译. 台北：宝鼎.

陈鸿基，严纪中. 2004. 管理信息系统[M]. 台北：双叶书廊.

胡世忠. 2013. 云端时代的杀手级应用：Big Data海量资料分析[M]. 台北：天下杂志.

简祯富. 2014a. 从台积电案例，谈大数据分析如何提升制造智能[M]//哈佛教你精通大数据. 台北：哈佛商业评论杂志.

简祯富. 2014b. 决策分析与管理：紫式决策分析以全面提升决策质量[M]. 台北：双叶书廊.

简祯富，施义成，林振铭，等. 2005. 半导体制造技术与管理[M]. 新竹：清华大学出版社.

中文论文与文章

IBM蓝色观点. 2011. 商业分析实用化 全球企业趋之若鹜[M]. IBM蓝色观点，39：16-17.

KITTLER R，WANG W. 1999. 数据分析渐露头角[M]. 中文半导体技术杂志：79-85.

Ou J. 2010. IBM BAO及对客户在商业分析与决策优化方面的价值[R]. IBM 2010新锐洞察高峰论坛.

简祯富，王兴仁，陈丽妃. 2005. 利用数据挖矿提升半导体厂制造技术员人力资源管理质量[J]. 品质学报，12(1)：9-28.

简祯富，李培瑞，彭诚涌. 2003. 半导体制程数据特征萃取与数据挖矿之研究[J]. 资讯管理学报，10(1)：63-84.

简祯富，林国胜. 2006. 建构cDNA生物芯片之二元数据挖矿模式及其实证研究[J]. 资讯管理学报，13(4)：133-159.

简祯富，林鼎浩，徐绍钟，等. 2001. 建构半导体晶圆允收测试资料挖矿架构及其实证研究[J]. 工业工程学刊，18(4)：37-48.

简祯富，林鼎浩，刘巧雯，等. 2002. 建构晶圆图分类之资料挖矿方法及其实证研究[J]. 工业工程学刊，19(2)：23-38.

简祯富，萧礼明，王兴仁. 2004. 建构半导体制造管理目标层级架构与制造数据之数据挖矿[J]. 工业工程学刊，21(4)：313-327.

彭金堂，张盛鸿，简祯富，等. 2005. 建构关联规则数据挖矿架构及其在台电配电事故定位之研究[J]. 资讯管理学报，12(4)：121-142.

王鸿儒，简祯富，李培瑞，等. 2002. 决策树资料挖矿架构及其于半导体制程之实证研究[J]. 科技管理学刊，7(1)：137-160.

谢佩原. 2011-5-13. 访清华大学简祯富教授：掌握转型机会 开创电子制造业新局[N]. 电子时报.

英文书籍

AGRESTI A. 1990. Categorical Data Analysis[M]. New York：John Wiley.

BASS P I, BASS F M. 2001. Diffusion of Technology Generations: A Model of Adoption and Repeat Sales [M]. Frisco, Texas: Bass Economics Inc.

BERGER J O. 1985. Statistical Decision Theory and Bayesian Analysis[M]. 2nd ed. New York: Springer-Verlag.

BERRY M, LINOFF G. 1997. Data Mining Techniques for Marketing, Sales and Customer Support[M]. New York: John Wiley and Sons.

BERRY M, LINOFF G. 2004. Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management[M]. 2nd ed. Indianapolis: Wiley.

BERSON A, SMITH S, THEARLING K. 2000. Building Data Mining Applications for CRM[M]. New York: McGraw-Hill.

BOX G E P, JENKINS G M. 1976. Time Series Analysis: Forecasting and Control[M]. San Francisco: Holden-Day.

BREIMAN L, FRIEDMAN J H, OLSHEN R A, et al. 1984. Classification and Regression Trees[M]. New York: Chapman & Hall.

BROCKWELL P J, DAVIS R A. 1991. Time Series: Theory and Methods[M]. New York: Springer-Verlag.

BROCKWELL P J, DAVIS R A. 2003. Introduction to Time Series and Forecasting[M]. 2nd ed. New York: Springer-Verlag.

CABENA P, HADJINIAN P, STADLER R, et al. 1997. Discovering Data Mining: From Concept to Implementation[M]. Upper Saddle River, New Jersey: Prentice Hall.

CRYER J D, CHAN K S. 2008. Time Series Analysis: with Applications in R[M]. 2nd ed. New York: Springer.

DRAPER N R, SMITH H. 1981. Applied Regression Analysis[M]. 2nd ed. New York: John Wiley & Sons.

DUDA R, GASCHNIG J, HART P E. 1979. Model design in the prospector consultant system for mineral exploration[M]//MICHIE D, Ed. Expert Systems for the Microelectronic Age. Edinburgh: Edinburgh University Press, 153-167.

FOX J, WEISBERG S. 2011. An R Companion to Applied Regression[M]. 2nd ed. Thousand Oaks CA: Sage.

FREEMAN J A, SKAPURA D M. 1991. Neural Networks: Algorithms, Applications, and Programming Techniques[M]. Reading, MA: Addison-Wesley.

GANTZ J, REINSEL D. 2012. Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East [M]. IDC iView, IDC report.

GURNEY K. 1997. An Introduction to Neural Networks[M]. London: UCL Press.

HAIR J F, BLACK W C, BABIN B J, et al. 2010. Multivariate Data Analysis: A Global Perspective[M]. 7th ed. Upper Saddle River, New Jersey: Pearson Prentice Hall.

HAN J, KAMBER M. 2006. Data Mining: Concepts and Techniques[M]. 2nd ed. New York: Morgan Kaufmann.

HAN J, KAMBER M, PEI J. 2011. Data Mining: Concepts and Techniques[M]. Waltham Morgan Kaufmann.

HASSOUN M H. 1995. Fundamentals of Artificial Neural Networks[M]. Cambridge: MIT Press.

HASTIE T, TIBSHIRANI R, FRIEDMAN J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. 2nd ed. New York: Springer.

IBM Corporation. 2011a. IBM Business Analytics and Optimization Software[M]. IBM Software Group.

IBM Corporation. 2011b. Business Analytics and Optimization: The New Competitive Edge[M]. White Paper.

JOHNSON R A, WICHERN D W. 2007. Applied Multivariate Statistical Analysis[M]. 6th ed. Englewood Cliffs, New Jersey: Prentice Hall.

KANTARDZIC M. 2003. Data Mining: Concepts, Models, Methods, and Algorithms[M]. Hoboken, New Jersey: Wiley-Interscience: IEEE Press.

KAUFMAN L, ROUSSEEUVW J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis [M]. New York: John Wiley & Sons.

KERR R. 1991. Knowledge-Based Manufacturing Management[M]. Sydney: Addison-Wesley.

KOHONEN T. 1995. Self-Organizing Maps[M]. Berlin: Springer-Verlag.

LaVALLE S. 2009. Business Analytics and Optimization for the Intelligent Enterprise[M]. New York: IBM Institute for Business Value, IBM Global Services.

LaVALLE S, HOPKINS M, Lesser E, et al. 2010. Analytics: The New Path to Value[M]. New York: IBM Institute for Business Value.

LEWIS C D. 1982. Industrial and Business Forecasting Method[M]. London: Butterworth.

LIU H, MOTODA H. 1998. Feature Selection for Knowledge Discovery and Data Mining[M]. Boston: Kluwer.

McNEIL D R. 1977. Interactive Data Analysis[M]. New York: Wiley.

PATTENSON D W. 1996. Artificial Neural Networks: Theory and Applications[M]. Singapore: Prentice Hall.

PAWLAK Z. 1991. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Boston: Kluwer Academic Publishers.

PINHEIRO J C, BATES D M. 2000. Mixed-Effects Models in S and S-PLUS[M]. Springer.

PYLE D. 1999. Data Preparation for Data Mining[M]. San Francisco, CA: Morgan Kaufmann.

QUINLAN J R. 1983. Learning efficient classification procedures and their application to chess end games [M]//MICHALSKI R S, CARBONELL J G, MITCHELL T M, Eds. Machine Learning: An Artificial Intelligence Approach. Los Altos: Morgan Kaufmann.

QUINLAN J R. 1993. C4. 5: Programs for Machine Learning[M]. San Francisco, CA: Morgan Kaufmann.

RAO V B, RAO H V. 1995. C++ Neural Networks and Fuzzy Logic[M]. New York: MIS Press.

RIPLEY B. 1996. Pattern Recognition and Neural Networks[M]. Cambridge: Cambridge University Press.

RUMELHART D E, McCLELLAND J L, the PDP Research Group. 1986. Parallel Distributed Processing [M]. Cambridge: MIT Press.

SCHROECK M, SHOCKLEY R, SMART J, et al. 2012. Analytics: The Real-World Use of Big Data [M]. New York: IBM Institute for Business Value.

SHARMA S. 1996. Applied Multivariate Techniques[M]. New York: John Wiley & Sons.

VANDAELE W. 1978. Participation in illegitimate activities: Ehrlich revisited, in deterrence and incapacitation[M]//BLUMSTEIN A, COHEN J, NAGIN D, Eds. Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. Washington DC: US National Academy of Sciences, 270-335.

VAPNIK V. 1995. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag.

VENABLES W N, RIPLEY B D. 2002. Modern Applied Statistics with S[M]. 4th ed. New York: Springer.

WHITE T. 2010. Hadoop: The Definitive Guide, Second Edition[M]. O'Reilly Media, Inc., Yahoo Press.

英文论文

AGRAWAL R, IMIELINKSI T, SWAMI A. 1993a. Mining association rules between sets of items in large databases[C]//Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington DC, USA.

AGRAWAL R, IMIELINKSI T, SWAMI A. 1993b. Database mining: A performance perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 5(6): 914-925.

AGRAWAL R, SRIKANT R. 1994. Fast algorithms for mining association rules[C]//Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile.

AGRAWAL R, SRIKANT R. 1996. Mining quantitative association rules in large relational tables[C]//Proceedings of the ACM-SIGMOD 1996 Conference on Management of Data, Montreal, Canada.

AKAIKE H. 1969. Fitting autoregressive models for prediction[J]. Annals of the Institute of Statistical Mathematics, 21(1): 243-247.

AKAIKE H. 1974. A new look at the statistical model identification[J]. IEEE Transactions on Automatic Control, 19(6): 716-723.

AKAIKE H. 1978. A Bayesian analysis of the minimum AIC procedure[J]. Annals of the Institute of Statistical Mathematics, 30(1): 9-14.

AMARI S. 1990. Mathematical foundations of neurocomputing[J]. Proceedings of the IEEE, 78(9): 1443-1463.

ANSLEY C F. 1979. An algorithm for the exact likelihood of a mixed autoregressive-moving average process[J]. Biometrika, 66(1): 59-65.

BARTLETT M S. 1937. Properties of sufficiency and statistical tests[C]//Proceedings of the Royal Statistical Society, Series A, 160(901): 268-282.

BARTO A G, Sutton R S, Anderson C W. 1983. Neuronlike adaptive elements that can solve difficult learning control problems[J]. IEEE Transactions on System, SMC-13(5): 834-846.

BASS F M. 1969. A New product growth model for consumer durables[J]. Management Science, 15(5): 215-227.

BASS F M, KRISHNANA T V, JAIN D C. 1994. Why the Bass model fits without decision variables[J]. Marketing Science, 13(3): 203-223.

BERGMEIR C, BENITEZ J M. 2012. Neural networks in R using the stuttgart neural network simulator: RSNNS[J]. Journal of Statistical Software, 46(7): 1-26.

BIXBY R E. 2002. Solving real-world linear programs: A decade and more of progress[J]. Operations Research, 50(1): 3-15.

BLENKO M W, MANKINS M C, ROGERS P. 2010. The decision-driven organization[J]. Harvard Business Review, 88(6): 54-62.

BOX G E P, Cox D R. 1964. An analysis of transformation[J]. Journal of the Royal Statistical Society, Series B, 26(2): 211-246.

BOX G E P, PIERCE D A. 1970. Distribution of the residual autocorrelations in autoregressive-integratedmoving average time series models[J]. Journal of American Statistics Association, 65(2): 1509-1526.

BREIMAN L. 1996. Bagging predictors[J]. Machine Learning, 26(2): 123-140.

BREIMAN L. 2001. Random forests[J]. Machine Learning, 45(1): 5-32.

CARPENTER G A, GROSSBERG S. 1987a. A massively parallel architecture for a self-organizing neural pattern Stephen recognition machine[J]. Computer Vision: Graphics and Image Processing, 37(1): 54-115.

CARPENTER G A, GROSSBERG S. 1987b. ART2: Self-organization of stable category recognition codes for analog input patterns[J]. Applied Optics, 26(33): 4919-4930.

CARPENTER G A, GROSSBERG S. 1990. ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures[J]. Neural Networks, 3(2): 129-152.

CARPENTER G A, GROSSBERG S, ROSEN D B. 1991. ART2-A: An Adaptive resonance algorithm for rapid category learning and recognition[J]. Neural Networks, 4(4): 493-504.

CATLETT J. 1991. On changing continuous attributes into ordered discrete attributes[C]//Proceedings of the European Working Session on Learning, 482: 164-178.

CERQUIDES J, de MANTARAS R L. 1997. Proposal and empirical comparison of a parallelizable distance-based discretization method [C]//Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'97), Newport Beach, California, 139-142.

CHIEN L F, Chien C-F. 2011. Manufacturing intelligence for class prediction and rule generation to support human capital decisions for high-tech industries [J]. Flexible Services and Manufacturing Journal, 23(3): 263-289.

CHIEN C-F. 2005. Modifying the inconsistency of Bayesian networks and a comparison study for fault location on electricity distribution feeders[J]. International Journal of Operational Research, 1(1/2): 188-203.

CHIEN C-F, CHANG K, CHEN C. 2003. Design of sampling strategy for measuring and compensating overlay errors in semiconductor manufacturing. International Journal of Production Research, 41(11): 2547-2561.

CHIEN C-F, CHEN L F. 2007. Using rough set theory to recruit and retain high-potential talents for semiconductor manufacturing [J]. IEEE Transactions on Semiconductor Manufacturing, 20 (4): 528-541.

CHIEN C-F, CHEN L F. 2008. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry[J]. Expert Systems with Applications, 34(1): 280-290.

CHIEN C-F, CHEN S L, Lin Y S. 2002. Using Bayesian network for fault location on distribution feeder [J]. IEEE Transactions on Power Delivery, 17(3): 785-793.

CHIEN C-F, CHEN Y J, Peng J T. 2010. Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle[J]. International Journal of Production Economics, 128(2): 496-509.

CHIEN C-F, HSU C-Y. 2006. A novel method for determining machine subgroups and backups with an empirical study for semiconductor manufacturing [J]. Journal of Intelligent Manufacturing, 17 (4): 429-440.

CHIEN C-F, HSU C-Y. 2014. Data mining for optimizing IC feature designs to enhance overall wafer effectiveness[J]. IEEE Transactions on Semiconductor Manufacturing, 27(1): 71-82.

CHIEN C-F, HSU C-Y, CHEN P L. 2013. Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence[J]. Flexible Services and Manufacturing Journal, 25(3): 367-388.

CHIEN C-F, HSU S C, CHEN Y J. 2013. A system for online detection and classification of wafer bin map defect patterns for manufacturing intelligence[J]. International Journal of Production Research, 51(8): 2324-2338.

CHIEN C-F, HSU C-Y, HSIAO C. 2012. Manufacturing intelligence to forecast and reduce semiconductor cycle time[J]. Journal of Intelligent Manufacturing, 23(6): 2281-2294.

CHIEN C-F, HSU C-Y, TSENG H Y. 2010. Manufacturing intelligence for semiconductor long-term demand forecast[C]//Proceedings of 11th Asia Pacific Industrial Engineering & Management Systems Conference, 07-10 December, Malaka, Malaysia.

CHIEN C-F, LIN K Y. 2012. Manufacturing intelligence for Hsinchu science park semiconductor sales prediction[J]. Journal of the Chinese Institute of Industrial Engineers, 29(2): 98-110.

CHIEN C-F, LIN K Y, YU A. 2014. User-experience of tablet operating system: An experimental investigation of Windows 8, iOS 6, and Android 4.2[J]. Computers and Industrial Engineering, 73: 75-84.

CHIEN C-F, WANG W C, CHENG J C. 2007. Data mining for yield enhancement in semiconductor manufacturing and an empirical study[J]. Expert Systems with Applications, 33(1): 192-198.

CHIEN C-F, WU J Z. 2003. Analyzing repair decisions in the site imbalance problem of semiconductor test machines[J]. IEEE Transactions on Semiconductor Manufacturing, 16(4): 704-711.

CHIEN C-F, WU C H, CHIANG Y S. 2012. Coordinated capacity migration and expansion planning for semiconductor manufacturing under demand uncertainties[J]. International Journal of Production Economics, 135(2): 860-869.

CHIEN C-F, ZHENG J N. 2012. Mini-max regret strategy for robust capacity expansion decisions in semiconductor manufacturing[J]. Journal of Intelligent Manufacturing, 23(6): 2151-2159.

CHOU W, CHIEN C-F, GEN M. 2014. A multiobjective hybrid genetic algorithm for TFT-LCD module assembly scheduling[J]. IEEE Transactions on Automation Science and Engineering, 11(3): 692-705.

CORTES C, VAPNIK V. 1995. Support-vector networks[J]. Machine Learning, 20(3): 273-297.

DARLIN D. 2006-7-1. Airfares made easy (or easier[N]). The New York Times.

DEAN J, GHEMAWAT S. 2004. MapReduce: Simplified data processing on large clusters[C]//Proceedings of the 6th USENIX Symposium on Operating Systems Design and Implementation (OSDI-04), 137-149.

DEAN J, GHEMAWAT S. 2008. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 51(1): 107-113.

DEMPSTER A P, LAIRD N M, Rubin D B. 1977. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B, 39(1): 1-38.

DUDA R O, HART P E, NILSSON N J. 1976. Subjective Bayesian Methods for rule-based inference systems[J]. National Computer Conference, 45: 1075-1082.

EHRLICH I. 1973. Participation in illegitimate activities: A theoretical and empirical investigation[J]. Journal of Political Economy, 81(3): 521-565.

ENGLE R F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation[J]. Econometrica, 50(4): 987-1008.

ESTER M, KRIEGEL H, SANDER J, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 226-231.

FARAWAY J, CHATFIELD C. 1998. Time series forecasting with neural networks: A comparative study

using the airline data[J]. Applied Statistics, 47(2): 231-250.

FAYYAD U, IRANI K. 1993. Multi-interval discretization of continuous-valued attributes for classification learning[C]//Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1022-1027.

FAYYAD U, PIATETSKY-SHAPIRO G, SMYTH P. 1996. The KDD process for extracting useful knowledge from volumes of data[J]. Communication of ACM, 39(11): 27-34.

FENG L, LU H, YU J, et al. 1999. Mining inter-transaction associations with templates[C]//Proceedings of the Eighth International Conference on Information and Knowledge Management, 225-233.

FRIEDMAN N, GEIGER D, GOLDSZMIT M. 1997. Bayesian network classifiers[J]. Machine Learning, 29(2-3): 131-163.

GHEMAWAT S, GOBIOFF H, LEUNG S. 2003. The Google file system[C]//Proceedings of ACM Symposium on Operating Systems Principles (OSOP-03), 29-43.

GRANGER C W J, NEWBOLD P. 1976. Forecasting transformed series[J]. Journal of the Royal Statistical Society, Series B, 38(2): 189-203.

GROSSBERG S. 1976. Adaptive pattern classification and universal recoding, Ⅰ: Parallel development and coding of neural feature detectors[J]. Biological Cybernetics, 23(3): 121-134.

GROSSBERG S. 1987. Competitive learning: From interactive activation to adaptive response[J]. Cognitive Science, 11(1): 23-63.

GUHA S, RASTOGI R, SHIM K. 1998. CURE: An efficient clustering algorithm for large databases [C]//Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, 73-84.

HAN J, FU Y. 1995. Discovery of multiple-level association rule from large databases[C]//Proceedings of the International Conference on Very Large Databases, 420-431.

HAN J, PEI J, YIN Y. 2000. Mining frequent patterns without candidate generation[C]//Proceedings of 2000 ACM-SIGMOD International Conference Management of Data, 29(2): 1-12.

HO K M, SCOTT P D. 1997. Zeta: A Global method for discretization of continuous variables[C]//Proceedings of KDD97: 3rd International Conference of Knowledge Discovery and Data Mining, Newport Beach, CA, 191-194.

HO T K. 1998. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8): 832-844.

HOLTE R C. 1993. Very simple classification rules perform well on most commonly used datasets[J]. Machine Learning, 11: 63-91.

HSU C-Y, CHIEN C-F, LIN K Y, et al. 2010. Data mining for yield enhancement in TFT-LCD manufacturing and an empirical study[J]. Journal of the Chinese Institute of Industrial Engineers, 27(2): 140-156.

HSU S C, CHIEN C-F. 2007. Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing[J]. International Journal of Production Economics, 107(1): 88-103.

HUANG Y C. 2013. Mining association rules between abnormal health examination results and outpatient medical records[J]. Health Information Management Journal, 42(2): 23-31.

HUANG Z. 1998. Extensions to the K-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery, 2(3): 283-304.

HYNDMAN R J, KHANDAKAR Y. 2008. Automatic time series forecasting: The forecast package for R

[J]. Journal of Statistical Software, 27(3): 1-22.
ISLAM T, MEADE N. 1997. The diffusion of successive generation of a technology: A more general model[J]. Technological Forecasting and Social Change, 56(1): 49-60.
KAISER H F. 1970. A second generation little jiffy[J]. Psychometrika, 35(4): 401-415.
KAISER H F, Rice J. 1974. Little jiffy mark Ⅳ[J]. Educational and Psychological Measurement, 34(1): 111-117.
KASS G. 1980. An exploratory for investigating large quantities of categorical data[J]. Applied Statistics, 29(2): 119-127.
KLEISSNER C. 1998. Data mining for the enterprise[C]//IEEE Proceedings 31st Annual Hawaii International Conference on System Sciences, 17, 295-304.
KOHONEN T. 1982. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics, 43(1): 59-69.
KOMOROWSKI J, PAWLAK Z, POLWSKI L, et al. 1999. Rough sets: A tutorial[M]//PAL S K, SKOWRON A, Eds. Rough Fuzzy Hybridization, A New Trend in Decision Making. Singapore: Springer, 3-98.
KRAAIJVELD M A, MAO J, JAIN A K. 1995. A nonlinear projection method based on Kohonen's topology preserving maps[J]. IEEE Transactions on Neural Networks, 6(3): 548-559.
KUO C J, CHIEN C-F, CHEN C D. 2011. Manufacturing intelligence to exploit the value of production and tool data to reduce cycle time[J]. IEEE Transactions on Automation Science and Engineering, 8(1): 103-111.
KUSIAK A. 2001. Rough set theory: A data mining tool for semiconductor manufacturing[J]. IEEE Transactions on Electronics Packaging Manufacturing, 24(1): 44-50.
LIEVENS F, VAN DAM K, Anderson N. 2002. Recent trends and challenges in personnel selection[J]. Personnel Review, 31(5-6): 580-601.
LIN K S, CHIEN C-F. 2009. Cluster analysis of genome-wide expression data for feature extraction[J]. Expert Systems with Applications, 36(2): 3327-3335.
LIU B, HSU W, MA Y. 1999. Mining association rules with multiple minimum supports[C]//Proceedings of 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 337-341.
LIU C W, CHIEN C-F. 2013. An intelligent system for wafer bin map defect diagnosis: an empirical study for semiconductor manufacturing[J]. Engineering Applications of Artificial Intelligence, 26(5-6): 1479-1486.
LIU H, HUSSAIN F, TAN C L, et al. 2002. Discretization: An enabling technique[J]. Data Mining and Knowledge Discovery, 6(4): 393-423.
LU H, HAN J, FENG L. 1998. Stock movement and N-dimensional inter-transaction association rules[C]//Proceedings of 1998 SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery, Seattle, Washington, 12, 1-7.
NARENDRA P M, FUKUNAGA K. 1977. A branch and bound algorithm for feature subset selection[J]. IEEE Transactions on Computers, C-26(9): 917-922.
NORTON J A, BASS F M. 1987. A diffusion theory model of adoption and substitution for successive generations of high-technology products[J]. Management Science, 33(9): 1069-1086.
OOVEREEM A, ROBINSON J C R, Leijnse H, et al. 2013. Crowdsourcing urban air temperatures from smartphone battery temperatures[J]. Geophysical Research Letters, 40(15): 4081-4085.
PANG B, LEE L. 2008. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information

Retrieval, 2(1-2): 1-135.
PARK J S, CHEN M S, YU P S. 1995. An effective hash-based algorithm for mining association rules[J]. ACM SIGMOD Record Archive, 24(2): 175-186.
PAWLAK Z. 1982. Rough sets[J]. International Journal of Computer and Information Sciences, 11(5): 341-356.
PAWLAK Z. 1996. Why rough sets? [C]//Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, 2: 738-743.
PAWLAK Z. 1997. Rough set approach to knowledge-based decision support[J]. European Journal of Operational Research, 99(1): 48-57.
PEI J, HAN J. 2000. Can we push more constraints into frequent pattern mining? [C]//Proceedings of 2000 International Conference on Knowledge Discovery and Data Mining, Boston, MA.
PENG C, CHIEN C-F. 2003. Data value development to enhance yield and maintain competitive advantage for semiconductor manufacturing[J]. International Journal of Service Technology and Management, 4 (4-6): 365-383.
PENG J T, CHIEN C-F, TSENG T L B. 2004. Rough set theory for data mining for fault diagnosis on distribution feeder[J]. IEE Proceedings-Generation, Transmission, and Distributions, 151(6): 689-697.
POTVIN C, LECHOWICZ M J, TARDIF S. 1990. The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures[J]. Ecology, 71(4): 1389-1400.
QUINLAN J R. 1986. Introduction to decision trees[J]. Machine Learning, 1(1): 81-106.
SAVASERE A, OMIECINSKI E, Navathe S. 1995. An efficient algorithm for mining association rules in large databases[C]//Proceedings of the 21st International Conference on Very Large Data-Bases Zurich, 432-444.
SHAW M J, SUBRAMANIAM C, TAN G W. 2001. Knowledge management and data mining for market [J]. Decision Support Systems, 31(1): 127-137.
SMITH J W, EVERHART J E, DICKSON W C, et al. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus[C]//Proceedings of the Symposium on Computer Applications in Medical Care (Washington, 1988). Los Alamitos: IEEE Computer Society Press, CA, 261-265.
SMITH W R. 1956. Product differentiation and market segmentation as alternative marketing strategies [J]. Journal of Marketing, 21(1): 3-8.
SRIKANT R, AGRAWAL R. 1995. Mining generalized association rules[C]//Proceedings of International Conference on the Very Large Data Base, 407-419.
SCHWARZ G. 1978. Estimating the dimension of a model[J]. Annals of Statistics, 6(2): 461-464.
TAAM W, HAMADA M. 1993. Detect spatial effects from factorial experiments: an application from integrated-circuit manufacturing[J]. Technometrics, 35(2): 149-160.
THURAISINGHAM B. 2000. A primer for understanding and applying data mining[J]. IT Professional, 2 (1): 28-31.
TSENG T L B, JOTHISHANKAR M C, WU T. 2004. Quality control problem in printed circuit board manufacturing—an extended rough set theory approach[J]. Journal of Manufacturing Systems, 23(1): 56-72.
WALCZAK B, MASSART D L. 1999. Rough sets theory[J]. Chemometrics and Intelligent Laboratory Systems, 47(1): 1-16.
WARD J H. 1963. Hierarchical grouping to optimize an objective function[J]. Journal of the American

Statistical Association 58(301): 236-244.

WATSON H, WIXOM B. 2007. The current state of business intelligence[J]. IEEE Computer, 40(9): 96-99.

WEHRENS R, BUYDENS L M C. 2007. Self- and super-organizing maps in R: the Kohonen package[J]. Journal of Statistical Software, 21(5): 1-19.

ZEILEIS A, HOTHORN T. 2002. Diagnostic checking in regression relationships[J]. R News, 2(3): 7-10.

ZHANG T, RAMAKRISHAN R, LIVNY M. 1996. BIRCH: an efficient data clustering method for very large databases[C]//Proceedings of 1996 ACM SIGMOD International Conference on Management of Data, Montreal Canada, 103-114.

R 语言网络资源

BORTHAKUR D. 2008. HDFS architecture guide. The Apache Software Foundation[EB/OL]. [2015-06-07]. http://hadoop. apache. org/docs/r1. 2. 1/hdfs_design. html.

HAHSLER M, CHELLUBOINA S. 2014. arulesViz: Visualizing association rules and frequent itemsets, R package version 0. 1-8[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.

HAHSLER M, BUCHTA C, GRUEN B, et al. 2014. arules: Mining association rules and frequent itemsets, R package version 1. 1-1[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.

KUHN M, WESTON S, COULTER N. 2014, C50: C5. 0 decision trees and rule-based models, R package version 0. 1. 1-16[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.

LIAW A, WIENER M. 2014. randomForest: Breiman and Cutler's random forests for classification and regression, R package version 4. 6-10[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.

QUINLAN J R. 1998a. C5. 0: An informational tutorial[EB/OL]. [2015-06-07]. http://www. rulequest. com/see5-unix. html, RuleQuest Research.

QUINLAN J R. 1998b. Is See5/C5.0 better than C4.5? [EB/OL]. [2015-06-07]. http://rulequest. com/see5-comparison. html, RuleQuest Research.

RIPLEY B, VENABLES B, BATES D M, et al. 2014. MASS: Support functions and datasets for venables and Ripley's MASS, R package version 7. 3-30[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.

RIZA L S, JANUSZ A, CORNELIS C, et al. 2014. RoughSets: Data analysis using rough set and fuzzy rough set theories, R package version 1. 0-0[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.

SCUTARI M. 2014. belearn: Bayesian network structure learning, parameter learning and inference, R package version 3. 5[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.

The FoRt Student Project Team. 2013. CHAID: Chi-squared automatical interaction detection, R package version 0. 1-1[EB/OL]. [2015-06-07]. http://r-forge. r-project. org/projects/chaid/.

THERNEAU T, ATKINSON B, RIPLEY B. 2014. rpart: Recursive partitioning and regression trees, R package version 4. 1-5[EB/OL]. [2015-06-07]. http://CRAN. R-project. org/.